Integrated analysis of proteomics data to assess and improve the scope of mass spectrometry based genome annotation

Michael Mueller

Trinity Hall

A dissertation submitted to the University of Cambridge
for the degree of Doctor of Philosophy

European Molecular Biology Laboratory
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD
United Kingdom.

Email: mmueller@ebi.ac.uk

30 March 2009

To Daniela, Benjamin and Hannah

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university, and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

This dissertation does not exceed the specified length limit of 300 pages as defined by the Biology Degree Committee.

This dissertation has been typeset in 12 pt Palatino using LaTeX2ε according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

30 March 2009 Michael Mueller

Integrated analysis of proteomics data to assess and improve the scope of mass spectrometry based genome annotation

Abstract

Michael Mueller
Trinity Hall

30 March 2009

The completion of the human genome has shifted attention from deciphering the sequence to the identification and characterisation of the functional components. Availability of the genome sequence has fostered an array of high-throughput technologies to systematically probe gene function on a genome-wide scale at all levels of biological information flow, from the DNA sequence over transcripts to proteins. A powerful approach to study gene function on protein level is the identification and quantification of proteins in complex mixtures by mass spectrometry. Despite significant technological and methodological advances, the complexity and the dynamic range of proteomes still pose major challenges for the analysis of biological samples.

Based on an integrative bioinformatics analysis, I examine the composition of mass spectrometry proteomics datasets with respect to the coverage of the particular proteome under study as well as the protein-coding genome as a whole. Using the example of a large-scale collaborative study of protein expression in human brain tissue, I point out characteristics of mass spectrometry proteomics datasets in terms of resolution and functional composition. On the basis of a comprehensive survey of publicly available proteomics data in the context of the genome sequence, I assess to what extent the findings from the analysis of the brain proteome study reflect global trends in mass spectrometry based proteomics. Following on from the results obtained by the analysis of experimental data, I outline and evaluate a strategy to improve the selectivity and sensitivity of target driven proteomics through diversification of peptide populations by combinatorial proteolysis.

The results presented in this dissertation show that mass spectrometry is an indispensable tool for large-scale protein research. However, they also demonstrate profound shortcomings of the technology regarding composition, redundancy and resolution of the generated data, emphasising the need for more targeted and systematic approaches to proteomics. The proposed combinatorial strategy contributes towards this aim by significantly increasing the coverage of the proteome by protein-specific signature peptides that are suitable reporter candidates for targeted proteomics experiments based on single reaction monitoring.

Acknowledgements

This dissertation describes work carried out at the European Bioinformatics Institute (EBI) in Hinxton, UK, between April 2005 and March 2009. The EBI is is an outstation of the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. My research work at the EBI was funded through the EMBL international PhD programme.

I would like to thank my supervisor Rolf Apweiler for taking me on as a PhD student and for his support and guidance over the last four years. Many thanks also to the members of my thesis advisory committee Wolfgang Huber, Jyoti Choudhary, Sarah Teichman and Lars Steinmetz for their critical and constructive assessment of my work. I am equally thankful to Lennart Martens for his continuous interest in and support of my research and the very valuable feedback on my dissertation.

I also would like to thank the HUPO Proteome Proteome Project (BPP) bioinformatics committee for involving me in the bioinformatics analysis of the HUPO BPP pilot study.

The following people I would like to thank for accompanying me on this exciting journey through time and space at the EBI and in Cambridge.

Members, ex-members and visitors of the Proteomics Services Team, Lennart Martens, Juan Antonio Vizcaíno, Florian Reising, Matthieu Visser and Joe Foster, the lunch time deipnosophists and companions on the regular and at times extensive walks on and around the campus. Richard Côté, Phil Jones, Antony Quinn, Cathrine Leroy and Samuel Kerrien for their help with everything IT related and last but not least the team leader Henning Hermjakob.

Lennart, without you I might well never have made it to this stage. Thanks for sharing not only my research interests (or initiating them in the first place) but also a healthy dose of irony and cynicism. Conversations would only be half the fun without them. I am certain your enthusiasms is going to inspire many (PhD) students to come. Slow down a bit though, otherwise you will not only be a professor in five years but also one with grey hair (and probably no beard given your disapproval of beards). Juanan, you were a great colleague and it was always fun to work with you! Who will finish my chips at lunch time now? I hope you won't spot that some parts of this dissertation are "justified" and some are not. Florian, the only team member who really appreciates the saying "a busy life is a wasted life". I hope one day you will find yourself on the veranda of that beach bungalow in south america enjoying the sun. All it takes, is to convince Anna of the affore mentioned saying. Matthieu, thank you for always pointing out both sides of the coin not only in research but also life related matters. Soon, I will be joining you in the quest for the Holy Grail in the haystack of biopolymers. Yet, not bold enough to face the twenty-letter hydra I will humbly stick with the As, Cs, Gs and Ts for the time being. Richard, your brilliant linguistic idioms made me chuckle and the team meetings more enjoyable than meetings are generally meant to be. Phil, until recently the last British survivor in the all too European enclave of the PRIDE team. I don't know how you managed to stand all the moaning about single glasing and the lack of insulation over the years. Joe, you will have to hold the British end up from now on. Good luck with your PhD! And remember, never, never hand your private number to German professors. Antony, I would like to thank you for the good times we had together at the begining of my PhD at the group retreat, gigs and on various pub nights. Samuel, thank you for having been the only developer ever bold enough to let a layman touch code. Cathrine, thank you for extending my knowledge of French music. I still haven't managed to comb through all the 150 or so Serge Gainsbourg songs you once gave me. Henning, thank you for giving me a home in the vast spaces of the PANDA group.

I also would like to thank my fellow PhD students at the EBI and my friends in Cambridge for their support and major contribution to make life in Cambridge much more fun then I ever anticipated. In particular, I would like to thank Jörn, first and foremost for being a great friend and for the good times we had together. But also for persuading me that it is worthwhile spending an entire day trying to install an *R* package and for his helpful advice regarding this weird and wonderful super high-level "programming language". Jacky, thank you for resurrecting the social calendar after it had died a horrible death after the PhD course. I am now aware of at least a dozen music genres ending in "core" I had never even heard of before. John, I am grateful for your friendship and the countless evenings we spent together, philosophising about life, the universe and everything. And of course for proofreading this dissertation. I owe you one (or two or three). I also would like to thank Daniel, Melanie and Georg for the very enjoyable times we had together in particular as team mates during the many pubquizzes at the Clarendon. Unfortunately, I couldn't share the moment of glory with you. Furthermore, I would like to thank my long-lasting friends Tim and Thomas for the, well, long-lasting friendship even across long distances.

My greatest gratitude goes to my wonderful partner Daniela as well as my parents who always supported me unconditionally all these years. Daniela, thank you for taking me on this thrilling ride that we had over last ten years with all its ups and downs. Thank you for always sticking to me and for the encouragement and comfort in moments of dispear. Thank you for being the loving mother of our children Hannah and Benjamin who brought so much joy into our lifes. Finally, I am very grateful to my parents Else and Hans who always supported my decissions no matter how implausible they seemed to them. I thank you both for allowing me to go my way.

Contents

List of Figures

List of Tables

List of Acronyms & Abbreviations

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
AMT	accurate mass tag
API	application programming interface
BLAST	basic local alignment search tool
BPP	Brain Proteome Project
cDNA	complimentary DNA
CDS	coding sequence
CID	collision induced dissociation
CNS	central nervous system
COFRADIC	combined fractional diagonal chromatography
CSF	cerebrospinal fluid
cSNP	coding single nucleotide polymorphism
CVI	Cardiovascular Initiative
Da	Dalton
DAS	Distributed Annotation System
DCC	data collection centre
DDBJ	DNA Data Bank of Japan
ECDF	empirical cumulative distribution function
EMBL	European Molecular Biology Laboratory
ER	endoplasmic reticulum
ERGIC	endoplasmic reticulum-Golgi intermediate compartment
ESI	electrospray ionisation
EST	expressed sequence tag
FDR	false discovery rate

FT	Fourier transform
FTP	file transfer protocol
GFP	green fluorescent protein
GNF	Genomics Institute of the Novartis Research Foundation
GO	Gene Ontology
GOA	Gene Ontology Annotation
GPCR	G-protein coupled receptor
GPI	glycosylphosphatidylinositol
gpmDB	Global Proteome Machine Database
HAI	Human Antibody Initiative
hPDQP	human Proteome Detection and Quantitation Project
HGNC	Human Genome Organisation Genome Nomenclature Consortium
HGPI	Human Disease Glycomics/Proteome Initiative
HILIC	hydrophilic interaction liquid chromatography
HPA	Human Protein Atlas
HPI	Human Proteome Initiative
HPLC	high pressure liquid chromatography
HPP	Human Proteome Project
HPRD	Human Protein Reference Database
HTTP	hypertext transfer protocol
HUGO	Human Genome Organisation
HuPA	Human Proteinpedia
HUPO	Human Proteome Organisation
ICR	ion cyclotron resonance
IEC	ion exchange chromatography
IPI	International Protein Index
IT	ion trap
iTRAQ	isobaric tag for relative and absolute quantitation
KUPP	Kidney and Urine Proteome Project
LC	liquid chromatography
LIT	linear ion trap
LOPIT	localization of organelle proteins by isotope tagging
LPP	Liver Proteome Project

MALDI	matrix-assisted laser desorption/ionisation
MAPU	Max-Planck Unified Proteome Database
MHC	major histocompatibility complex
MMHD	Mouse Models of Human Disease Initiative
mRNA	messenger RNA
MS	mass spectrometry
MTLE	mesial temporal lobe epilepsy
MudPIT	multidimensional protein identification technology
NCBI	National Center for Biotechnology Information
NMR	nuclear magnetic resonance
nr	non-redundant
OMIM	Online Mendelian Inheritance in Man
PAGE	polyacryl amide gel electrophoresis
PIR	Protein Information Resource
PMF	peptide mass fingerprinting
PPP	Plasma Proteome Project
PRIDE	Protein Identification Database
PSI	Proteomics Standard Initiative
PST	peptide sequence tag
PTM	post-translational modification
Q	quadrupole
RPC	reverse phase chromatography
SDS	sodium dodecyl sulfate
SELDI	surface-enhanced laser desorption/ionisation
SLD	soft laser ionisation
SNP	single nucleotide polymorphism
SRM	selected reaction monitoring
TCA	tricarboxylic acid
TOF	time-of-flight
UniProtKB	UniProt Knowledgebase
URL	universal resource locator
UTR	untranslated region
UV	ultra violet
VEGA	Vertebrate Genome Annotation
XML	extensible markup language

Glossary

Minimal and maximal explanatory set of protein identifications In this text *minimal explanatory set* of protein identifications refers to the parsimonious or least complex set of proteins that can explain the peptide sequences observed in a mass spectrometry experiment. Often, for a given set of peptides many, equally valid minimal explanatory protein sets exist that cannot be distinguished based on the peptide evidence. *Maximal explanatory set* refers to the set of proteins that includes all proteins, the observed set of peptide sequences can potentially originate from. That is, the maximal explanatory set of protein identifications comprises all alternative hypothesis for the peptide evidence. For a given set of peptides there is only one maximal explanatory set of protein identifications.

Chapter 1

Introduction

1.1 Proteomics: the large-scale study of protein function

The Oxford Dictionary of Biochemistry and Molecular Biology defines the *proteome* as “the complete expression profile of the proteins of an organism” and *proteomics* as “the study of the proteome by the analysis of protein structure and composition” [1].

In general the term proteome, first coined by Marc Wilkins at the 1st Siena 2D-Electrophoresis meeting in 1994 [2], is loosely defined. It can refer to the protein complement not only of an organism, but also that of a subcellular structure, compartment, organelle, cell, or tissue ,expressed at a specific point in time, usually in a defined biological or pathological state. The idea of identifying and analysing all proteins encoded by a genome was proposed in the mid 1990s [3] and the first proteome-scale analysis of a eukaryote was presented by Shevchenko and colleagues [4] in 1996.

1.2 Annotating the human proteome

Any genome is only as valuable as its annotation [5], and the immense task of adding value to the human genome sequence has only just begun. Functional sequence elements, including genes need to be identified and their function elucidated. The classical reductionist approach to function determination is no longer sufficient due to the magnitude of the undertaking. Instead, an array of 'omics' technologies have emerged to probe the function of many or all genes and their products in a single experiment. The field of proteomics has produced tools to systematically study proteins, enabling their identification in complex protein mixtures, the study of protein-protein interactions, the determination of protein structure and subcellular localisation, the measurement of changes in their abundance, distribution and modification as well as the characterisation of their biochemical activity.

1.2.1 The magnitude of the task

When the first draft of the human genome sequence was published in 2001 [6, 7] the number of protein-coding genes was estimated to be around 30000 to 40000. With the availability of the nearly finished sequence of the human euchromatic genome [8], along with expanded complimentary DNA (cDNA) libraries, genomes of other organisms and improved computational methods for gene prediction this number has been adjusted to 20000 to 25000 protein-coding genes. The Ensembl automatic genome annotation system [9] reports 21804 protein-coding genes in the current release of Ensembl (53) [10, 11], which is based on the National Center for Biotechnology Information (NCBI) assembly 36 (November 2005). Of these 21731 are known and 73 'novel' genes. The former are predicted on the basis of curated and partially curated species-specific evidence while predictions of the latter are based on other experimental evidence such as protein and cDNA sequence information from related species.

Automated gene annotation has now been complemented by manual curation, including a careful review of gene structure and an examination of expressed sequence tag (EST) and transcript evidence. The Vertebrate Genome Annotation (VEGA) database is a central repository for manually annotated genome sequences which currently covers 20 of the 24 human chromosomes [12]. Although greatly improving the gene models, manual annotation has not significantly affected the total gene count.

The relatively low number of human genes suggests that the complexity of human biology is achieved through regulation on the transcriptional, post-transcriptional and post-translational level. Alternative splicing [13, 14] and translation [15, 16] as well as post-translational modifications such as phosphorylation, glycosylation and proteolytic processing contribute to a “proteomic stratification” that produces a protein population with a diversity that is several orders of magnitude higher than that of the number of genes encoding them. Correspondingly, it has been estimated that the human proteome comprises as many as 1000000 different protein species [17].

1.2.2 The study of gene function at the protein level

Studying gene function at the protein level is vital to the understanding of how cells, tissues and organisms perform their function. Proteins are the active agents of a cell. They are the end-point of biological information flow from the genome sequence over messenger RNA (mRNA) to the gene product and thus closest to the phenotypic manifestation of a genotype. Their assembly into functional modules such as protein complexes, compartments and pathways can be studied directly only at the protein level. Furthermore, proteins are subjected to a high degree of post-translational modification and regulation. Variations in protein isoforms and quantities that underlie phenotypic variations are often difficult or impossible to deduce from DNA or transcript level information alone.

Functional studies of proteins can be divided into three categories: i) protein identification, ii) protein characterisation and iii) protein quantification. Protein *identification* experiments provide evidence for the presence of a protein in a sample and are used to identify components of protein complexes or the protein constituents of a subcellular compartment, cell or tissue. Protein *characterisation* studies aim to determine the biochemical and biophysical properties of proteins. This includes measurement of enzymatic activities, determination of the three-dimensional structures of proteins and complexes, measurement of affinity to other biopolymers or studies of transient and stable modification of proteins. The goal of protein *quantification* is the determination of the stoichiometry of proteins in complexes, organelles, pathways and cells and its variation across different physiological and pathophysiological states.

1.2.3 From single protein characterisation to global approaches

The classical reductionist approach to studing complex biological systems is to dissect them into individual components and determine the connections between the components on the basis of their individual physicochemical properties [18]. Though this approach has been extremely successful in elucidating many biological phenomena over the last century insights gained from the genome sequence gave the impulse for a paradigm shift away from the reductionist viewpoint towards a system level approach to biology. It has been recognised that specificity of complex biological activity is not only achieved through specificity at component level but is also the result of the complex interactions between the components [18]. While the availability of the sequenced genome forms an important part of this reasoning, it is also the basis for an expansion of molecular biology from single molecule to genome-wide analyses [19]. Along with recent technological advances, it has enabled experimental approaches in proteomics capable of measuring protein expression, post-translational modifications and subcellular localisation of proteins as well as their assembly into complexes and pathways on a proteome-wide scale.

1.3 Identification of proteins in complex samples

Central to many proteomics approaches is the ability to identify proteins in complex samples. Proteins can be identified based on structural and compositional features. Sequence based approaches rely on information about the primary structure, i.e. the ordered amino acid composition of the protein. Immunological approaches identify a protein using antibodies that specifically recognise and bind features of the tertiary, i.e. three-dimensional, structure of the protein.

1.3.1 Immunological identification

Immunological identification is based on the detection of proteins by antibodies, highly specific affinity agents, produced by the immune system to detect and eliminate exogenous molecules [20].

Advantages of antibody based protein identification are the high sensitivity and specificity of antibodies. Antibodies can detect very low quantities of protein in many different contexts *in situ* and *ex situ*. This makes them very versatile tools for protein identification.

The availability of genome sequences has enabled the systematic large-scale production of antibodies. Bioinformatics algorithms are employed to identify coding sequences suitable for high-throughput cloning and expression of protein fragments. These fragments are subsequently used as antigens to raise antibodies against the target protein [21].

Despite significant automation, however, antibody generation and validation is still a laborious and expensive process. Currently, antibodies are only available for a relatively limited set of proteins. Once antibodies become available for a larger percentage of proteins, protein arrays - ordered sets of antibody probes - will become an important tool for the large-scale measurement of protein expression. So-called inverse phase arrays are already used for the systematic analysis of tissue expression on protein level [22]. This technology has recently been extended to include cell samples [23].

1.3.2 Sequence based identification

Sequence based identification of proteins is based on the determination of their composition; protein sequencing refers to the determination of the ordered amino acid composition of a protein. Technological, methodological and computational limitations currently do not allow to determination of the sequence of a protein directly. Instead the sequence of a subset of its peptides is determined which is then used for protein identification.

1.3.2.1 Protein identification by Edman degradation

Until the early 1990s the most common approach used to directly determine the amino acid sequence of a peptide was Edman degradation. The method, introduced by Per Edman in 1950, is based on cyclic degradation of proteins with phenylisothiocyanate [24]. The detached amino acids are subsequently identified by ultra violet (UV) absorbance spectroscopy. Despite the automation of the process in the late sixties [25, 26, 27], with a cycle time of one hour per amino acid, determination by Edman degradation remains a slow and inefficient process.

1.3.2.2 Protein identification by mass spectrometry

In the mid 1990s Edman degradation was replaced by mass spectrometry (MS) based approaches as the method of choice to determine the amino acid sequences of peptides. In MS the elemental composition of a sample is determined based on the mass-to-charge (m/z) ratio of ionised molecules or molecule fragments generated from the sample. The m/z ratio is measured based on the motion of the ions as they pass through an electromagnetic field [28].

The foundations of modern mass spectrometry were laid by Eugen Goldstein and Wilhelm Wien in the late 19th century. The first fully functional mass spectrometers were built by Arthur Demster and Francis Aston in 1918 and 1919 [29].

Mass spectrometry of biopolymers including proteins and peptides had become feasible towards the end of the 1980s due to the introduction of two new ionisation methods, electrospray ionisation (ESI) [30] and soft laser ionisation (SLD) more commonly referred to as matrix-assisted laser desorption/ionisation (MALDI) [31, 32]. These so-called 'soft' ionisation methods enabled the ionisation of large, non-volatile biomolecules too fragile for ionisation by more conventional methods, making them amenable to analysis by a mass spectrometer [33, 34]. Mass spectrometry is central to the rapid developments in the field of proteomics over the last two decades as it has enabled the high-throughput identification of proteins in complex mixtures. It is a primary and secondary tool in many experimental approaches in proteomics some of which will be introduces later on in this chapter. First, mass spectrometry based protein identification will be discussed in detail in the following.

1.4 The mass spectrometry based proteomics experiment

An overview of a typical MS based proteomics experiment is shown in figure 1.1. Proteins are prepared from a biological specimen, e.g. a tissue sample or cell culture. The protein extract may then be fractionated for example by gel electrophoresis or centrifugation. For practical reasons, which will be discussed in more detail below, it is peptides, not intact proteins, that are subjected to mass spectrometry. Proteins are digested with a sequence specific protease, usually trypsin, and the resulting peptides are separated by chromatography, e.g. high pressure liquid chromatography (HPLC) or ion exchange chromatography (IEC). Peptides leaving the chromatography column are ionised for injection into the mass spectrometer. This happens either online in the case of ESI, which can be directly coupled to the chromatograph (as depicted in figure 1.1) or offline in the case of MALDI where liquid chromatography (LC) fractions are first deposited on a metal target before automated analysis. The most common mode of mass spectrometry is tandem MS, a two stage process, that results in two different types of spectra being recorded: firstly, the *MS*, *MS1* or *precursor ion* spectrum, which is the mass spectrum of peptides that eluted from the column at a specific point in time and secondly, the *MS/MS*, *MS2* or *product ion* spectrum, which is the mass spectrum of fragment ions generated from selected precursor ions, usually through collision with an inert gas. The final step is the computational analysis to obtain peptide sequence information from the recorded spectra by matching them against theoretical spectra derived from a sequence database. A protein sequence database is then searched with the obtained peptide sequence information for protein inference. In the following sections the individual steps of a MS experiment are discussed in more detail.

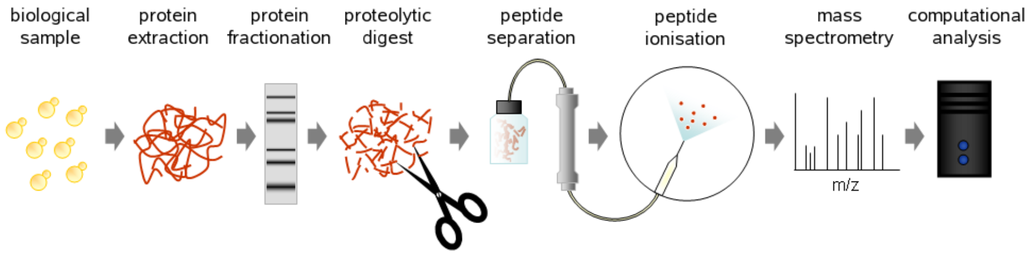


Figure 1.1: Overview of a typical MS based proteomics experiment. Starting from a biological sample, proteins are extracted, which may or may not be fractionated, for example by SDS-PAGE, before they are subjected to a proteolytic digest, usually with trypsin. The resulting peptides are separated, often by HPLC, and ionised for analysis in a mass spectrometer. The recorded mass spectra are evaluated computationally for peptide sequencing and protein identification.

1.4.1 Protein preparation and fractionation

Depending on the sample preparation and fractionation steps, proteomics techniques can be divided into *gel-based* and *gel-free* methods.

The first approach to characterise complex protein mixtures was separation by two-dimensional (2D)-PAGE followed by MALDI-MS [35]. Proteins are separated by their isoelectric point in the first dimension and their mass in the second dimension. The advantage of gel-based methods is the low degree of complexity of the analysed gel spots as well as the straightforward resolution of protein isoforms. However, the method suffers from various shortcomings such as the limited dynamic range, throughput and sample coverage. Another problem is the selection against hydrophobic proteins inherent to gel-based separation, hampering the analysis of important protein classes such as transmembrane proteins [36, 37, 38]. Therefore, this approach is best suited to analyse less complex protein samples or characterise specific proteins in detail.

Since the beginning of this decade, efforts have been focused on the development of gel-free techniques allowing a less biased and more comprehensive analysis of complex protein samples. In an approach also referred to as 'shotgun' proteomics, peptides obtained from the digestion of unseparated proteins are extensively fractionated by capillary chromatography and analysed by automated MS analysis. Online coupling of chromatographic separation methods with ESI makes these techniques particularly suitable for high-throughput analysis of protein mixtures. While conventional HPLC can be used to analyse peptide mixtures of lower complexity [39], high-resolution methods based on multi-dimensional fractionation are used for the analysis of highly complex samples [40]. Advantages of shotgun proteomics are increased proteome coverage and scalability. Disadvantages, on the other hand, are the high complexity of the analysed peptide mixture and the large amount of spectral data generated, which poses significant informatics challenges regarding peptide and protein identification. Furthermore, shotgun proteomics suffers from a limited dynamic range and a high degree of redundancy [41].

Different approaches have been taken to address the issues arising from sample complexity. Sample complexity can be reduced by focusing on the analysis of subcellular fractions [42, 43, 44], isolated organelles [45, 46] or other subproteomes such as phosphorylated or glycosylated proteins [47], instead of complete cell lysates or tissue extracts. Other approaches are based on the depletion of highly abundant proteins mostly by immunoaffinity methods [48, 49] or reducing differences in protein concentration using so-called 'equaliser' technolgy [50].

1.4.2 Proteolytic digest

Although technically possible, intact proteins are rarely analysed in proteomics. Instead protein samples are usually analysed by a 'bottom-up' approach in which proteins are broken down proteolytically into peptides which are identified and then mapped back to the protein sequence.

There are several reasons for and advantages to analysing peptides instead of proteins. Firstly, the performance of mass spectrometers is much better in typical peptide mass ranges of up to 4000 Dalton (Da). In addition, the interpretation of mass spectra obtained from intact proteins is computationally difficult as it is practically impossible to predict the mass of a mature and potentially modified protein from primary sequence information. Therefore, sequence information is indispensable for reliable protein identification. This is inferred from fragmentation spectra, which are best generated and interpreted from short peptides rather then proteins. Additionally, proteins are more difficult to handle then peptides and breaking proteins down into peptides circumvents problems arising from the physicochemical properties of proteins affecting solubility or affinity [51].

Drawbacks of the bottom-up approach are the absence of information on the parent protein that a peptide was generated from and the challenges this poses to protein inference [52]. Furthermore, only a fraction of all peptides emitted by a protein are observable by mass spectrometry [53]. So while peptide information is often sufficient for protein identification it does not allow a full characterisation of a protein.

A pre-requisite for the application of a protease to generate peptides for a proteomics experiment is the lack of substrate specificity, i.e. the capability to cleave any protein inside its peptide chain. Proteases fulfilling this criterion are pancreatic and gastric endopeptidases like trypsin and pepsin. The serine protease trypsin is the most commonly used enzyme to produce peptides for mass spectrometric analysis. This is for several reasons: firstly, trypsin is a highly stable and efficient protease with high sequence specificity [54, 55]. Secondly, trypsin cleaves proteins carboxy (C)-terminal of arginine and lysine residues, thereby generating peptides in a mass range and with fragmentation properties beneficial for MS, resulting in mass spectra of high information content whose interpretation is relatively uncomplicated [51].

Other proteases are used to create peptide populations complementary to trypsin, impacting on detection as well as sequence coverage and specificity [56, 57, 58, 59, 60]. Less complex populations of longer peptides, for example, can be generated using proteases with only one cleavage site, like the serine protease Lys-C which cleaves C-terminal of lysine and the cysteine protease Arg-C which cleaves C-terminal of arginine [61, 62]. While Lys-C is a highly specific protease, the specificity of Arg-C is less predictable [63]. Also of lower specificity as well as decreased activity is pepsin A, an aspartate protease which cleaves predominantly C-terminal to phenylalanine and leucine [64]. Another protease that finds application in proteomics and is relevant in the context of this work is the *Staphylococcus aureus* serine proteases V8 [65, 60]. In digestion buffers that contain phosphate the protease cleaves after glutamic and aspartic acid (V8DE). In the absence of phosphate the protease is specific to glutamic acid (V8E) residues. Proteases relevant in the context of this work are listed in table 1.1.

1.4.3 Peptide separation

Prior to injection into the mass spectrometer, proteolytic peptides are separated by chromatography, usually HPLC. HPLC separates peptides by their hydrophobicity, based on their solubility in an organic solvent ver-

Table 1.1: Proteases used for peptide generation in mass spectrometry based proteomics. Listed are proteases commonly used in proteomics and relevant in context of this work. All listed proteases cleave C-terminal of the amino acid residue.

protease	enzyme class	cleavage sites
Trypsin	serine	lysine(K), arginine(R)
Arg-C	cysteine	arginine (R)
Lys-C	serine	lysine (K)
Pepsin A	aspartate	phenylalanine (F), leucine (L)
V8E	serine	glutamic acid (E)
V8DE	serine	glutamic acid (E), aspartic acid (D)

sus their affinity for a hydrophobic stationary phase [66].

Multi-dimensional methods are used to achieve a higher resolution of peptide separation. In so-called geLC-MS, protein mixtures are first separated by one-dimensional (1D)-SDS-PAGE, followed by an LC separation of proteins contained in individual gel slices [67]. Alternatively, two orthogonal chromatography steps can be combined to achieve a higher degree of separation. This is implemented in the popular multidimensional protein identification technology (MudPIT) approach that separates peptides by charge and hydrophobicity through a combination of strong cation exchange and reverse phase chromatography (RPC) [68, 40]. Separation by MudPIT can be improved further by implementing it on an ultrahigh-pressure scale [69]. More recently free-flow chromatography, separating peptides by their isoelectric point, as well as hydrophilic interaction liquid chromatography (HILIC) [70], have been coupled with RPC for proteomic analyses [37, 71, 72]. HILIC has been found to have advantages for the targeted analysis of post-translationally modified proteins [73].

Fractionation can also be achieved after the biochemical separation and peptide ionisation, in the mass spectrometer: gas phase fractionation separates peptides in the m/z dimension by iteratively analysing peptides separated by LC over multiple, defined m/z ranges [71, 74, 75].

1.4.4 Ionisation

ESI and MALDI, the two ionisation methods that paved the way for the application of MS to proteins, are still the most important approaches to ionise and transfer peptides into the gas phase.

To ionise a sample by MALDI it is embedded in a crystalline organic matrix which absorbs light of a specific wavelength, usually in the UV range. The sample is mixed with the matrix, dissolved in an organic solvent and spotted on a sample plate. After the solvent has evaporated and the matrix has crystallised, the analyte is sublimated in high vacuum from the matrix by a laser pulse and becomes ionised by receiving protons from the matrix molecules. MALDI generally results in peptide ions with a charge of $z = +1$ [76].

While MALDI is the more robust technique regarding contamination of the sample by salts or detergents [77] a drawback is that, although possible [78], online coupling of MALDI to LC is cumbersome. This limits the scalability of the technology. Nevertheless, MALDI remains a valuable ionisation technique which is applied mainly to the analysis of less complex samples and to complement results obtained by ESI [79, 80].

Online coupling of ESI to capillary based chromatographic separation is straightforward. Peptides eluting from the chromatography column are sprayed through a capillary or metal needle into a strong electromagnetic field. This generates highly charged droplets which decrease in size through evaporation of the solvent. The decreasing droplet size results in an increase of the charge density and droplet fission. Peptide ions are generated either through desorption from the droplet surface or through continuous droplet fission, eventually leading to every droplet containing only a single analyte ion [76, 79]. Generally, ESI results in ions with a charge of $z = +2$ or higher.

A variation of the MALDI method, called surface-enhanced laser desorption/ionisation (SELDI) should also be mentioned in this context, as it is of relevance to mass spectrometry in proteomics. In SELDI the analyte is spotted on a surface modified with a chemical functionality. Affinity of certain proteins in the sample to the chemical group results in the binding of these proteins to the surface. After unbound proteins have been washed off, the matrix is allowed to crystallise and the sample is ionised like in MALDI [81].

1.4.5 Mass spectrometry

1.4.5.1 Mass analysers

The mass analyser is the centre piece of any mass spectrometry proteomics experiment. There are different types of mass analysers which all have in common the fact that they determine mass-to-charge ratios of analyte ions moving in an electromagnetic field. The four types of mass analysers important in proteomics are the ion trap (IT), time-of-flight (TOF), quadrupole (Q) and Fourier transform (FT)-ion cyclotron resonance (ICR) mass analysers which are briefly described in the following.

Ion Trap analyser In an ion trap mass analyser ions are accumulated over time by *trapping* them in a region of a vacuum system by electric fields. MS1 and MS2 analysis are both performed in the same unit.

The original three-dimensional (3D) or quadrupole ion trap invented in 1953 [82] consists of two hyperbolic end-cap electrodes and a central ring electrode. Ions of an m/z value of interest are captured at the focal point of the electromagnetic field between the three electrodes by adjusting the voltage accordingly, while ions of other m/z values are ejected. Upon fragmentation of the precursor ions through collision with an inert gas, product ions are ejected in order of their m/z value by iterating over a voltage range and the m/z values recorded by the detector. Although the 3D IT has high sensitivity, the small size limits the number of ions that can be trapped, resulting in low resolution and mass accuracy [79].

The 2D or linear IT was introduced in 2002 [83, 84]. It consists of four parallel electrodes arranged along a cylindrical space. Ions are trapped by creating a potential along the axis of the electrodes. The increased ion storage capacity of linear ITs allows for increased sensitivity, resolution and mass accuracy [80]. Consequently, linear ITs have largely replaced the traditional 3D instruments.

A relatively recent development is the orbitrap [85] which is based on a mass analyser originally invented by Makarov [86]. In this type of ion trap ions oscillate *orbitally* around the axis of an electrostatic field generated by an outer barrel-like electrode and a coaxial inner spindle-like electrode. Using Fourier transformation the m/z values of ions can be calculated based on the frequency with which they oscillate along the central spindle. Orbitraps combine a very high resolution with high mass accuracy [87].

Time-of-Flight analyser Time-of-flight (TOF) MS is based on the acceleration of ions along an electric field of known strength [88]. Ions of the same charge receive the same kinetic energy but travel at different speeds depending on the mass-to-charge ratio. By measuring the time the ion needs to cross a defined distance its m/z ratio can be calculated. TOF instruments are superior in mass resolution to ion traps. However, the resolution is inversely correlated with the ion mass. Another factor influencing resolution is the spreading of ions in three dimensional space after acceleration, resulting in not all ions of the same mass reaching the detector at the same time. Two strategies used to address this problem, delayed pulse extraction [89] and reflectron TOF [90], will not be discussed further here.

Quadrupole analyser The quadrupole analyser consists of four cylindrical rods aligned in parallel to each other, and functions as a mass filter [91]: ions pass an oscillating electrical field between the rods in a spiralling trajectory, with the radius of the spiral depending on the m/z ratio of the ion. By adjusting the voltage accordingly, only ions of a specific m/z ratio will traverse the field, while ions of other m/z ratios will collide with the rods. Quadrupoles are used as highly selective mass filters in applications where ions of interest need to be detected with high specificity, an approach which gains increasing importance in current MS proteomics [92, 93].

Fourier transform ion cyclotron resonance analyser In an FT-ICR mass analyser ions move in circles within a strong electromagnetic field generated by a cyclotron [94]. A cyclotron is a particle accelerator that accelerates ions by a combination of high-frequency alternating voltage and a magnetic field. The mass-to-charge ratio of the ions is determined based on their cyclotron frequency which is measured when the ions pass near a pair of detection plates. The signal consists of superimposed sine waves of an ion's frequency and intensity which are untangled by Fourier transformation to generate the mass spectrum [94]. FT-ICR analysers can determine masses at unprecedented resolution, accuracy and sensitivity and are the current state-of-the-art in mass spectrometry. However, due to the high instrument and operating costs the technology is not yet widely used in proteomics.

1.4.5.2 Mass spectrometers

The different types of mass analysers are used in various configurations, either on their own or combined in hybrid systems, to make up the mass spectrometer. Depending on the identification strategy, which will be discussed later, only one mass spectrum - that of the precursor ions - or two mass spectra - that of the precursor and product ions, will be recorded. Figure 1.2 shows a schematic of a generic mass spectrometer configuration.

The type of mass analyser used often correlates with the ionisation method. Although TOF instruments can be operated with ESI and MALDI, they are mostly coupled to MALDI ion sources. Ion traps, quadrupole and FT-ICR analysers are found in setups that generate ions by ESI.

MALDI-TOF instruments are used to determine the mass of intact peptide

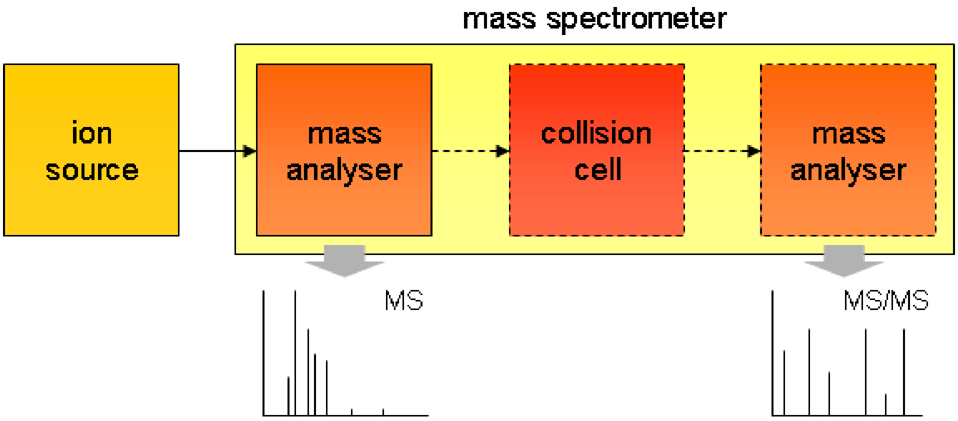


Figure 1.2: Schematic of a mass spectrometer. A mass spectrum (MS) of ions entering the mass spectrometer is recorded by the first mass analyser. In tandem MS, precursor ions of a specific m/z ratio are chosen for fragmentation in the collision cell. The second mass analyser then records the mass spectrum of the resulting product ions (MS/MS).

ions for peptide identification by peptide mass fingerprinting [95, 96] (see section 1.4.6.1). In TOF/TOF instruments [97] two consecutive TOF sections are separated by a collision cell. The precursor spectrum is recorded by the first TOF analyser which selects ions of a particular m/z ratio for fragmentation in the collision cell. The second analyser separates the resulting fragment ions and records the product ion spectrum. TOF analysers are also used in conjunction with a quadrupole [98, 99]. In QqTOF instruments the Q quadrupole serves as a mass filter to select ions for fragmentation by a second non-filtering *collision* quadrupole (q). Alternatively, the ions can be led through both quadrupoles for measurement of the whole mass spectrum in the TOF section. Quadrupole-TOF instruments are used with ESI and MALDI ion sources. They have a high sensitivity, resolution and accuracy [80] and are particularly well suited for quantitation [79].

Pure quadrupole (QqQ) instruments have the same configuration as QqTOF instruments except the TOF analyser is replaced by a third quadrupole. Ions may be either scanned or filtered in the third quadrupole. Triple quadrupole instruments are equally well suited for quantitation [79] .

Operating the third quadrupole as a linear ion trap results in a hybrid instrument increasingly important in proteomics due to its capability to perform selected reaction monitoring (SRM). That is, the occurrence of a particular precursor and product ion combination, also referred to as a *transition*, can be selectively detected by using both quadrupoles as mass filters. Furthermore, QqIT instruments are capable of precursor ion as well as neutral loss scanning and therefore are frequently applied to the analysis of protein modifications [83, 100].

Recently FT-ICR mass analysers have been combined with linear ion traps [101] which has enabled the real-time recording of tandem mass spectra at high resolving power, mass accuracy, and dynamic range [102, 103].

1.4.6 Data analysis

There are various identification strategies based on different types of MS experiments, which require different methods of data analysis. In the following some important identification strategies are discussed, of which tandem MS is currently the most frequently used approach in proteomics.

1.4.6.1 Identification strategies

Peptide mass fingerprinting peptide mass fingerprinting (PMF) is based on the assumption that a protein can be uniquely identified based on a combination of protein mass and the masses of a set of peptides generated by digesting the protein with a protease. A non-MS based PMF method was first proposed in the late 1970s by Laemmli and colleagues who used the fragment pattern of proteolytic peptides separated by gel electrophoresis to identify proteins isolated from an SDS gel [104].

The classic MS based method uses peptides of proteins isolated from a 2D gel analysed by MALDI-TOF [35, 95, 105, 96]. A computer algorithm which takes a protein mass range and peptide masses as input is then used to search a database of theoretical peptide masses to identify the protein [35]. Higher sequence coverage and improved protein identification can be achieved by identifying 2D gel spots by LC-MS [106].

While this approach is reliable for smaller proteomes peptide mass redundancy and insufficient mass accuracy increase the possibility of false positives for more complex protein samples. Post-translational protein modifications altering the protein and peptide masses further complicate the approach.

Accurate mass and time tag In shotgun proteomics the effect of ambiguities arising from peptide mass redundancy and low mass accuracy is amplified by the absence of information about the parent protein mass. The advent of high-accuracy MS for proteomics led to the proposal of a strategy to identify proteins based solely on peptide mass that is applicable to shotgun proteomics data. The accurate mass tag (AMT) strategy presented by Smith *et al.* in 2000 assumes that a protein sequence can be identified by standard LC-MS with high confidence based solely on the mass of a single peptide if measured with sufficient precision [107]. It was later demonstrated that the approach could successfully identify AMTs for more then 60% of the proteins encoded in the genome of the procaryote *Deinococcus radiodurans* using FT-ICR mass spectrometry [108, 109]. However, *Deinococcus radiodurans* is a bacterium with a relatively small number of around 3200 protein-coding genes [110]. With increasing complexity of the proteome, the ability to identify proteins on the basis of peptide mass alone, even if measured with high accuracy, becomes increasingly difficult.

To improve performance of the AMT strategy an extension of the approach has been devised which combines peptide mass with LC retention time [110]. Critical to the success of the approach is the standardisation of elution times, i.e. the reproducibility of elution times across runs, which is complicated by column drift associated with temperature changes and flow rate [111, 112]. Various machine learning approaches have been applied to predict LC retention times and align multiple analyses [112]. Apart from the technical challenges, estimations of the specificity of the results suggest a limited performance of the approach for complex proteomes, even at high mass accuracy levels [111]. This, together with further limitations arising from the lack of prediction algorithms for modified peptides as well as non-standard LC separation techniques [112], has so far prevented the widespread application of this approach.

Tandem mass spectrometry The first step towards contemporary peptide identification strategies based on tandem MS spectra was made in 1994 with the peptide sequence tag (PST) approach. By complementing peptide mass with partial sequence information from the peptide start and end regions, obtained by fragmentation of the precursor peptide, the specificity of sequence database searches could be significantly increased [113]. An automated method to identify peptides by computational correlation of observed tandem mass spectra with theoretical spectra derived from an *in silico* digest of a sequence database was published the year after by Yates and colleagues [114]. Peptide identification by spectral matching is still the main approach to peptide identification in shotgun proteomics today. Methods to determine the peptide sequence *de novo*, directly from the mass differences in the fragmentation spectrum, are emerging [115] but due to the complexity of mass spectra are still beyond the scope of current knowledge and computational means [116, 117] and will not be discussed further here. Reference [118] provides a comprehensive review of current computational methods for protein identification including *de novo* sequencing approaches. The following section gives a brief overview of approaches based on spectral matching and some of the most frequently used algorithms.

1.4.6.2 Peptide identification by spectral matching

The principle of identifying peptides by spectral matching is implemented in a range of different search algorithms. Before these are discussed in more detail the main characteristics of product ion spectra are briefly outlined.

Fragmentation of a precursor peptide results in two different types of product ion series, derived from the carboxy (C)-terminal or the amino (N)-terminal end of the peptide. Furthermore, the peptide backbone can break at different positions depending on the experimental approach. In collision induced dissociation (CID) breakage mainly occurs at the amide bonds [119]. Figure 1.3 illustrates the different product ions that are produced from fragmentation of the precursor ion at the amide bonds and explains their nomenclature according to Biemann and Roepstorff [120, 121]. It also shows the location of the corresponding $y-$ and $b-$ion peaks in the spectrum. $Y-$ions are the predominant ions in CID spectra obtained by QqQ and QqTOF instruments. CID spectra acquired by IT instruments contain both a series of $y-$ and $b-$ions. Experimental spectrum peak intensities vary due to differences in the efficiency of bond breakage between different amino acids. Algorithms to predict peptide fragmentation have been devised [122, 123]. Furthermore, it should be noted that additional ion types can arise from loss of NH_3 and H_2O.

Experimental and theoretical spectra are correlated by either matching peaks directly or by comparing the integral intensities of corresponding spectrum intervals [76]. While the difficulty in the first approach lies in the identification of corresponding peaks, the second approach suffers from a loss of precision and problems arising from peaks coinciding with interval boundaries [76].

Algorithms for spectral matching can be divided into probabilistic and non-probabilist methods depending on the scoring scheme employed [76].

Probabilistic methods score the spectral correlation by calculating the probability that the matching spectrum originates from the matched candidate sequence. Various probabilistic approaches including Bayesian and odds ratio based methods are employed. MASCOT [124] and X!Tandem [125] are two prominent examples of probabilistic spectral matching algorithms. Other examples include SCOPE [126] and OLAV [127].

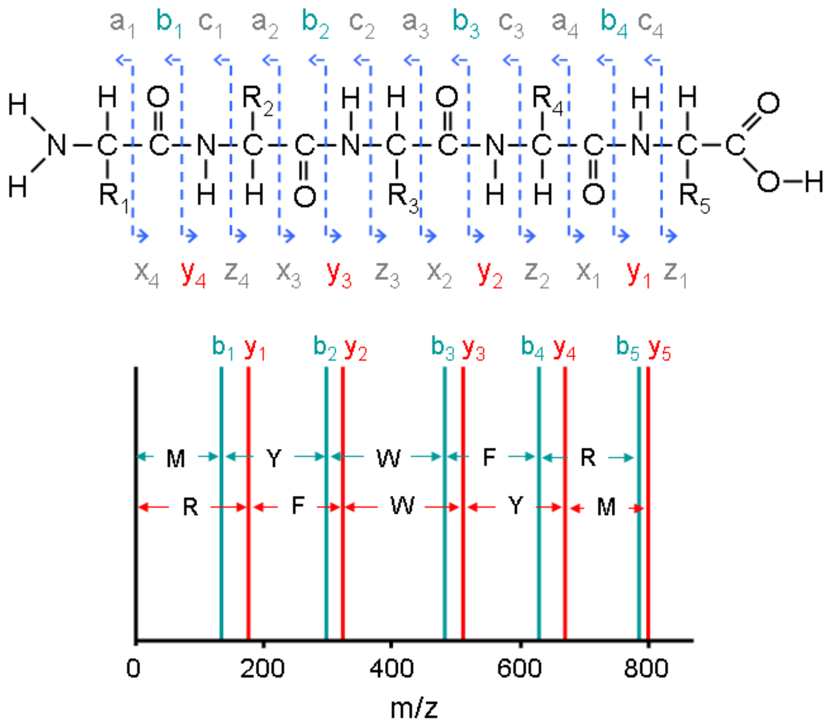

a1 b1 c1 a2 b2 c2 a3 b3 c3 a4 b4 c4
R1 R2 R3 R4 R5
x4 y4 z4 x3 y3 z3 x2 y2 z2 x1 y1 z1
b1 y1 b2 y2 b3 y3 b4 y4 b5 y5
M Y W F R
R F W Y M
0
200
400
600
800
m/z

Figure 1.3: The product ion spectrum. Shown is the Biemann and Roepstorff [120, 121] nomenclature of product ions which is based on the location of the peptide backbone breakage and whether the ion is derived from the N-terminal or C-terminal part of the precursor ion. Shown below are the theoretical locations of $y-$ and $b-$ion peaks in the fragmentation spectrum of the hypothetical peptide MYWFR. The y_5- and b_5-ions correspond to the entire peptide sequence.

In principle non-probabilistic scoring of spectrum matches is based on the number of matching spectrum peaks or intervals and their intensities. The SEQUEST^{TM} algorithm, developed as part of the work by Yates *et. al.* mentioned above, is an example of a non-probabilistic method which uses cross-correlation to score spectral matches [114].

Estimation of the error rate is a fundamental issue in peptide identification based on spectral matching. A common approach is the use of ′decoy′ databases which are shuffled or reversed versions of the sequence database used to generate the theoretical spectra [128]. The false positive rate is then evaluated based on the frequency of spectra matches against the decoy database. PeptideProphet is an algorithm that scores the accuracy of peptide identifications based on an empirical statistical model [129]. The model is trained on the characteristics (database search scores and number of tryptic termini) of peptide assignments known to be correct/incorrect and can then be used to classify peptide identifications of unknown confidence. An alternative approach is to increase the confidence of peptide and protein identifications by combining the results of several different search algorithms [130, 131, 132]. Identifications that the different methods converge on, are assigned a higher confidence.

1.4.6.3 Protein Inference

The final step in the analysis of mass spectrometry data is the inference of the precursor proteins for the observed set of proteolytic peptides. The reliability of protein identifications inferred from the available peptide evidence depends on two factors. Firstly, sequence similarity between homologous proteins, splice isoforms and conserved domains often results in ambiguities in the peptide-to-protein mapping and thus several alternative hypotheses that can explain the observed peptides [52].

There are two different strategies of dealing with the problem of peptide-to-protein ambiguity. One is to apply the principle of parsimony and determine the least complex set of proteins that can explain the peptide sequences observed in a mass spectrometry experiment. This set is commonly referred to as the *minimal explanatory set* of proteins. Often, for a given set of peptides many, equaly valid minimal explanatory protein sets exist that cannot be distinguished based on the peptide evidence. The alternative solution to the problem is to construct a *maximal explanatory set* of protein identifications that includes all proteins, the observed set of peptide sequences can potentially originate from, i.e. comprises all alternative hypothesis for the observed peptide evidence. For a given set of peptides there is only one maximal explanatory set of protein identifications.

The second factor that influences the reliability of protein identification is the error rate of peptide identification. As protein identifications are assembled from peptide sequences the false positive rate in the peptide identification step is propagated to the level of protein identification [133]. Various approaches to controle the false discovery rate on the level of individual experiments have been developed [133, 134, 135, 136, 137, 138]. More recently, a strategy enabling the assessment of the confidence of protein identifications across different experiments and heterogeneous data sets, independent of the inference method and data set size has been published [133].

1.5 Proteomics approaches to probe protein function on a large scale

Having discussed the methodological and technical aspects of mass spectrometry, we now turn to the experimental approaches employed in proteomics to probe protein function on a large scale. While many of these approaches use mass spectrometry as a primary or secondary tool, there are also a range of non-MS based methods. For completeness some of these non-MS based methods will be mentioned as well.

1.5.1 Alternative splicing

It is estimated that around 50% of human multi-exon genes are alternatively spliced. Alternative splicing modulates subcellular localisation and activity of proteins through the insertion or deletion of functional domains, subcellular sorting signals or transmembrane regions [139, 140]. Perturbation of alternative splicing is implicated in a range of diseases [141].

Traditionally splice events are delineated by the analysis of nucleotide sequence data like EST and full length mRNA sequences [142, 143, 144]. More recently mircroarray technology and short read sequencing approaches have been employed for the high-throughput analysis of alternative splicing [145, 146].

Proteomics technologies are also increasingly used to extend genome annotations [147, 148]. Resources such as PeptideAtlas annotate the protein-coding genome with identified peptides, confirming predicted splice sites and isoforms [149]. Some groups have used the entire human genome, translated in all six reading frames, as a search base for tandem mass spectra, effectively detecting peptide evidence for novel protein-coding genes and gene models [150]. Although it has been suggested that proteomics data should be made an integral part of the genome annotation process to improve gene models, the required infrastructure is still missing [151]. One of the caveats of shotgun proteomics, with regards to the task of resolving splice isoforms, is the inherent difficulty of assigning peptides to isoforms; a result of sequence similarity between isoforms. Only a minority of peptides generated from alternative products of the same gene are isoform-specific [152].

1.5.2 Post-translational modification

In addition to alternative splicing, permanent or transient modification of proteins during or after translation contributes to the generation of a complex proteome from a limited genome. Post-translational modifications such as phosphorylation, proteolytic cleavage or glycosylation affect various aspects of protein behaviour such as activity, turnover, localisation and molecular interactions.

1.5.2.1 Phosphorylation

Phosphorylation is a reversible post-translational modification (PTM) involved in signalling cascades, the regulation of enzyme activity and the modulation of molecular interactions. There are a number of experimental approaches to map phosphorylation in protein mixtures. Methods based on 2D gel electrophoresis detect phosphorylated proteins by anti-phospho-amino-acid antibodies, metabolic labelling or phosphatase treatment. Protein spots are subsequently identified by mass spectrometry. Alternatively, phosphorylated proteins can be identified in complex mixtures by LC-MS either with or without prior affinity purification [153, 154]. In the latter case, fragmentation spectra that could not be assigned to proteins in a database of unmodified sequences are used in a second search iteration allowing for modification. Another approach used to map phosphorylation sites is the proteolysis of proteins by phospho-specific cleavage, followed by mass spectrometry. The observed cleavage pattern subsequently allows the deduction of phosphorylation sites [155]. However, unexpected behaviour of proteolytic enzymes such as transpeptidation [156] can obfuscate some of the sites in these approaches. An emerging technology that is likely to play an important role in phosphorylation profiling is protein array analysis [157, 158], though it should be noted that this method necessitates prior knowledge of the targets. Phospho-epitopes can also be studied using flow cytometry, yielding an activation-state readout of the intracellular environment [159].

As information about the dynamics of protein phosphorylation is often more informative than simply expanding the phosphoproteome "parts list", quantitative mass spectrometry methods have also been incorporated into phosphoproteomics [160, 161, 162, 163].

1.5.2.2 Proteolytic cleavage

Proteolytic cleavage is an important non-reversible PTM controlling the fate of proteins by influencing their subcellular localisation and activity. The 'degradome' is defined as firstly, the complete set of proteases expressed at a specific time by a cell, tissue or organism and secondly, the substrate repertoire of a protease [164]. The human genome encodes 561 proteases and protease-homologs [165]. It has become evident that proteolysis is a highly selective and regulated process playing a role in cellular processes that goes beyond non-specific protein catabolism [166]. For example, proteolysis can give rise to proteins, so-called cryptic neo-proteins with functions different from the parent proteins they were derived from.

At present it is not possible to predict peptidase cleavage sites computationally as substrate structure plays an important role in substrate specificity. Several proteomics studies based on diverse approaches, including iTRAQ (isobaric tag for relative and absolute quantitation) labelling [167], 2D gel electrophoresis followed by tandem MS and N-terminal COFRADIC (combined fractional diagonal chromatography) [168] have been dedicated to the discovery of new protease targets. These have lead to the identification of new potential metalloproteinase [169, 170] and caspase substrates [171, 172].

1.5.2.3 Glycosylation

Post-translational modification through transfer of glycans or carbohydrates to proteins is a complex process requiring the concerted action of a series of glycosyl transferases, each catalysing a specific step in the pathway. The conservation of this complicated process throughout evolution suggests that important functions are attached to protein glycosylation. Unfortunately, these functions remain poorly understood.

Nearly half of all known proteins are potentially glycosylated [173] and oligosaccharides are implicated in a multitude of crucial processes: regulation of protein folding, endoplasmic reticulum (ER)-to-Golgi transport, cell-cell communication, immune response and tumour biology [174, 175, 176, 177, 178]. Due to the non-template driven nature of glycosylation, it is not possible to knock out specific structures to evaluate their effect on the phenotype. A direct study of the functions of glycosylation is therefore difficult.

A number of different technologies have been employed to study glycosylation, including mass spectrometry [179, 180, 181], nuclear magnetic resonance (NMR) [182] and liquid chromatography for "glycan sequencing" [183, 184]. Glycan microarrays are used for glycan-protein interaction profiling [185] and new techniques to fluorescently label glycans have enabled quantification of glycan species on the array [186]. The carbohydrate binding specificity of lectins is exploited to measure protein glycosylation states using lectin microarrays [187, 188, 189].

1.5.3 Protein-protein interaction

Biological systems are modular, that is components with diverse molecular functions are dynamically assembled into complexes and pathways that accomplish a particular task [190]. Therefore, to understand biological processes at the system level it is essential to elucidate the complex network of interactions between the molecular components of a cell, including proteins [191].

Genome scale protein interaction maps based on the yeast-two-hybrid system [192] - ectopic expression of bait and prey proteins in the yeast nucleus and activation of a reporter gene upon their interaction - have been available for several model organisms for some time [193, 194, 195, 196].

An alternative strategy to the yeast-two-hybrid approach is based on affinity tagging of bait proteins and purification of protein complexes by co-immunoprecipitation followed by detection of complex components by MS [197]. Large-scale human interaction maps based on this approach have recently become available [198, 199, 200]. However, these cover only a subset of the proteome. Besides these large-scale datasets, previous studies have investigated interactions in the context of specific biological processes such as TNF-α/NF-κ B signal transduction [201], the heat shock protein (Hsp) 90 interactome [202] and epidermal growth factor (EGF) signalling [203].

1.5.4 Subcellular localisation

Eukaryotic cells, especially mammalian cells, are highly compartmentalised. The function of a protein is therefore often strongly correlated with its localisation. Bioinformatics analysis of sequence features such as signal peptides, transit sequences, nuclear localisation sequences, transmembrane regions or glycosylphosphatidylinositol (GPI) anchor sequences can be used to predict where in a cell a protein potentially resides [204, 205, 206, 207, 208, 209, 210].

Experimental procedures to determine subcellular localisation of proteins include the identification of protein constituents of isolated organelles by MS, or microscopy based analysis of localisation *in vivo* using fluorescently labelled versions of proteins. In mammalian cells large-scale localisation studies are complicated by splice isoforms and the wealth of different cell types.

The development of methods to systematically tag proteins with green fluorescent protein (GFP) using full-length cDNA libraries [211], in combination with microscopy-based high-throughput visual screening [212], has enabled analysis of protein localisation on a genome-wide scale. As part of a large-scale project to systematically analyse protein function, the German Cancer Research Centre, in collaboration with EMBL Heidelberg, have determined the localisation of $\sim$2000 fluorescently labelled proteins encoded by full-length cDNAs generated by the German cDNA Consortium, focusing on novel proteins with completely unknown function [213, 214].

Classical subcellular fractionation methods from cell biology, followed by gel-based and gel-free mass spectrometry approaches are applied in organelle proteomics to catalogue and quantify the protein constituents of cellular compartments and components [215]. Several proteomics studies have analysed human organelles and other cellular structures including nucleus [216, 217], nuclear membrane [218], endoplasmic reticulum-Golgi intermediate compartment (ERGIC) [219], Golgi [220], exosomes [221], centrosome [222], mitotic spindle [223] and mitochondria [224].

Often it is difficult to isolate pure organelles by fractionation. Furthermore, organelle organisation is not static and many proteins shuttle between compartments by membrane trafficking. Organelle proteomics strategies that take dynamic changes of protein composition into account, e.g. the localization of organelle proteins by isotope tagging (LOPIT) approach, are based on quantitative mass spectrometry analysis of subcellular fractions of cell lysates separated by density gradient centrifugation to determine organelle distribution profiles of proteins [225, 226, 227]. A recent analysis of a lymphocyte cell line by LOPIT to assign proteins to the

ER, Golgi, lysosomes, mitochondria and the plasma membrane indicates a highly dynamic organisation of organelles [228].

1.5.5 Tissue expression

While comprehensive information on tissue expression is available at the transcript level [229], establishment of a similar resource for protein expression is much more challenging. A large-scale project to systematically raise antibodies against each and every human protein with the aim of cataloguing tissue expression at the protein level is now underway.

The Human Protein Atlas (HPA) project [230], a collaboration between Uppsala University and the Royal Institute of Technology, has been established to interrogate tissue expression on protein level using antibody based proteomics [22]. The project systematically raises antibodies against human proteins, which are then used in high-throughput immuno-histochemistry screens on tissue arrays for expression profiling [231, 232]. Images are electronically captured and annotated. Currently, expression information for more than 1900 proteins in 48 different normal tissue types and 20 different types of cancer is accessible through the HPA web site (February 2009). Added value from such a project will come from the availability of validated antibodies and protein clones with great potential in research, diagnostics and therapeutics.

Recently several MS based proteomics studies to catalogue protein constituents of tissues and bodily fluids have been conducted. A collaborative proteomics study of human plasma has generated a core dataset of 3020 proteins identifications [233]. This figure has recently been further refined by taking into account multiple hypothesis testing [234]. Other human bodily fluids that have been profiled by MS include saliva [235] and cerebrospinal fluid [236]. A study of protein expression in human brain tissue - which is also the subject of this dissertation - has yielded a high confidence set of 1804 proteins in human temporal lobe specimens [132].

1.5.6 Targeted proteomics approaches

In order to overcome some of the shortcomings of shotgun proteomics such as the high degree of redundancy of protein identifications, limited reproducibility and the difficulty in resolving low abundance proteins, target-driven proteomics approaches are enjoying growing popularity as an alternative to global "discovery" approaches [41].

Instead of an unbiased profiling of all proteins in a sample, targeted approaches focus on a limited set of proteins of interest important to test a given hypothesis [41]. These sets are compiled based on existing biological knowledge about the system under study, obtained, for example, from the literature or other 'omics' platforms like microarrays.

A technology commonly used for this type of experiment is selected reaction monitoring (SRM), a highly selective and sensitive mode of tandem mass spectrometry which allows for the targeted detection of specific peptides. SRM is used to monitor the occurrence of proteotypic signature peptides, i.e. peptides observable by mass spectrometry that uniquely identify the targeted protein [53, 237] (see figure 1.4).

SRM is typically performed on a triple quadrupole mass spectrometer which cycles through a series of precursor and product ion m/z combinations or *transitions* which are characteristic for the monitored peptides, and records the signal [79]. Filtering at MS1 and MS2 level not only results in highly selective detection of peptides but also high sensitivity because of i) a decreased background noise due to the small population of precursor ions transmitted by the first mass analyser, and ii) an increase in the duty cycle of the mass spectrometer due to the non-scanning mode of operation [79]. The duty cycle is the fraction of ions with a certain m/z that are effectively analysed by the mass spectrometer [28].

1.5.7 Towards a human proteome project

The idea of systematically cataloguing all human proteins, including information on their expression, modifications and subcellular localisation, was

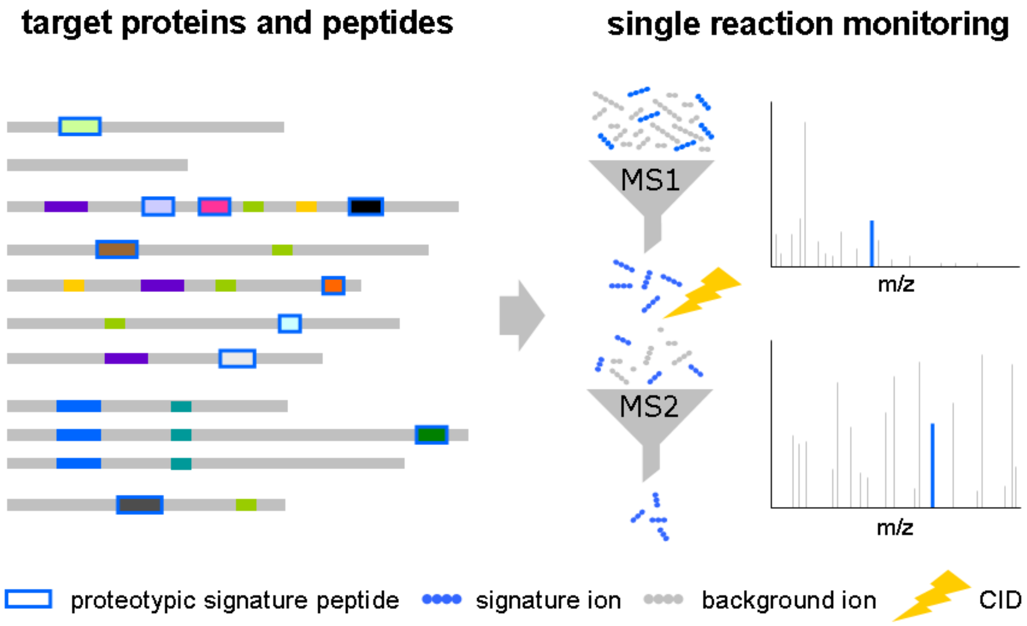

target proteins and peptides
single reaction monitoring
MS1
MS2
m/z
m/z
proteotypic signature peptide
signature ion
background ion
CID

Figure 1.4: Targeted proteomics based on selected reaction monitoring. A list of target proteins is compiled for which a set of proteotypic signature peptides is identified that can be used as surrogates to selectively monitor the respective protein (left side of the figure). Grey lines represent protein sequences, coloured stretches represent proteotypic peptides. The same colour indicates identical peptide sequences. Note that not all protein sequences have a proteotypic peptide. Blue boxes highlight proteotypic signature peptides which are observable and whose sequence is unique across the proteome. Proteotypic signature peptides are selectively detected by selected reaction monitoring (right side of the figure), an MS mode where the first mass analyser filters precursor ions of a specified mass and the second mass analyser monitors the occurrence of a product ion produced by CID which is characteristic for the targeted signature peptide.

first mooted in the early 1980s by Norman G. Anderson [238]. When the topic was raised again by Marc Wilkinson in the mid 1990's [3] technological advances, and in particular the advent of mass spectrometry in protein sciences, had brought the proteomics community significantly closer to achieving this ambitious goal. The idea gained momentum again in 2001 after the publication of the human genome sequence, which for the first time allowed a truly systematic study of all human protein products [239].

However, it was clear from the beginning that the challenges involved in a systematic study of the human proteome would by far exceed that of the human genome sequence. Unlike the Human Genome Project, which relied on one robust technology for DNA sequencing, proteomics uses an array of technologies with varying performance, and historically no coordination of world wide proteomics efforts comparable to that of the human genome consortium. The latter issue had started to be addressed shortly beforehand by the foundation of the Human Proteome Organisation [240] which will be discussed below.

Despite ongoing efforts, eight years on a Human Proteome Project (HPP) has still not been formalised. The view that a systematic mapping of proteins is a worthwhile effort to leverage information from the genome sequence and form the basis for future studies, however, is now more widely accepted in the proteomics community [241]. This view is emphasised by the fact that there is still no experimental evidence on protein level for the expression of around 44% of human protein-coding genes [241].

The reasons for the HPP evolving so slowly are technical, financial and organisational. Discussions about the conceptual and methodological approaches to take are still ongoing. The current plans envisage a 10 year project at an estimated cost of $1 billion, which maps at least one representative protein isoform per gene using a mixture of mass spectrometry and antibody based approaches [241, 242]. A scaled-down project, the human Proteome Detection and Quantitation Project (hPDQP), which could function as a pilot study for the HPP was proposed recently. The hPDQP would focus on the establishment of a catalogue of (non-isoform) specific peptide reporters for every protein-coding gene complemented by peptide-specific antibodies with an emphasis on the development of assays for biomarker detection [243].

1.6 The Human Proteome Organisation

Soon after the large-scale study of the human proteome had started to take off, the proteomics community realised that the magnitude of the task would require a high degree of coordination and collaboration. In analogy to the Human Genome Organisation (HUGO) that coordinated the analysis of the human genome, the Human Proteome Organisation (HUPO) was founded in 2001 with the aim of promoting international cooperation in the field of proteomics [240]. Since its foundation, eleven initiatives have been launched under HUPO's umbrella, dedicated to the study of selected tissues, organs, bodily fluids and stem cells, as well as the establishment of standards and resources enabling and facilitating collaborative protein analysis on a large scale.

The HUPO Liver Proteome Project (LPP) [244], the HUPO Brain Proteome Project (BPP) [245], the HUPO Plasma Proteome Project (PPP) [246], the HUPO Cardiovascular Initiative (CVI) [247], the Proteome Biology of Stem Cells Initiative [244] and HUPO Kidney and Urine Proteome Project (KUPP) [248] all aim to map the proteomes of the respective systems with a focus on the discovery of biomarkers for the development of diagnostics and therapeutics for human diseases. The Human Disease Glycomics/Proteome Initiative (HGPI) is focused on the sub-proteome of glycoproteins and aims to identify disease-related glycosylation changes in blood and urine. The Disease Biomarker Initiative is an integrative HUPO initiative that groups biomarker related projects in different organs [249]. The Mouse Models of Human Disease Initiative (MMHD) is still in the build-up phase. Many of the projects have initiated pilot projects to explore the portfolio of available proteomics techniques, establish standard operating procedures and set up the infrastructure for data collection, integration and dissemination.

While there is collaboration amongst the various HUPO projects, there are two initiatives of overarching relevance to all HUPO related activities as well as to the proteomics community as a whole. One is the Human Antibody Initiative (HAI) [250] which aims to generate a comprehensive catalogue of validated antibodies against every human protein within ten years. The aforementioned Human Protein Atlas was setup by the HAI. Apart from the antibodies systematically generated as part of the HPA project, already existing affinity reagents from research groups and commercial companies will also be incorporate into the catalogue. The second initiative of wider impact on proteomics research is the Proteomics Standard Initiative (PSI) [251]. The initiative was founded to address the need for systematic approaches to model, capture, and disseminate proteomics experimental data to facilitate data comparison, exchange and verification [252]. PSI has developed and implemented standards for the representation of mass spectrometry [253] as well as protein interaction data [254].

1.7 Proteomics repositories and databases

1.7.1 Primary mass spectrometry proteomics data repositories

With the move from small-scale, single-lab studies towards large-scale, collaborative, high-throughput approaches, the proteomics community quickly acknowledged the need for resources to systematically harvest the large amounts of experimental data generated to facilitate sharing and comparison of data [252, 255]. Over the years a number of repositories have emerged that capture MS proteomics data and make it publicly available. They all differ in the way data is submitted and made accessible, their functionality and their content. Relevant in the context of this work are the Protein Identification Database (PRIDE) [256], the Global Proteome Machine Database (gpmDB) [257], PeptideAtlas [149] and the Human Proteinpedia (HuPA) [258]. For a comprehensive review of public proteomics repositories see [259, 260].

1.7.1.1 Global proteome machine database

The gpmDB, hosted by the University of Manitoba, Canada, was originally developed as a web front end for the X!Tandem search engine. Data is submitted in the form of mass spectra together with a limited amount of experimental metadata. Spectra are subsequently identified mainly based on protein sequence information from the Ensembl database. The submitter can choose if the data is only processed or also stored in the database. Identifications can be accessed on peptide level as well as in the context of the protein sequence, where the identification frequencies of peptides originating from different parts of the protein are indicated. Libraries of confidently identified proteotypic peptides for various species can be downloaded.

1.7.1.2 PeptideAtlas

The Institute of Systems Biology in Seattle, USA, has developed the resource as a pipeline for genome annotation with proteomics data. PeptideAtlas accepts submissions of peptide identifications and peak lists, along with a minimal amount of information about the experiment. Submitted peak lists are processed by the Trans Proteomic Pipeline which consists of the SEQUEST^{TM} search engine and the PeptideProphet validation algorithm for identification [261]. Identified peptides are integrated with the genome sequence by mapping them to Ensembl protein sequences using the basic local alignment search tool (BLAST) [262]. Similar to gpmDB, PeptideAtlas visualises peptides in the context of the protein sequence, together with information about observability and peptide uniqueness across the genome. Proteotypic peptide sets are also available for download.

1.7.1.3 Protein identification database

Developed at the European Bioinformatics Institute, Hinxton, UK, PRIDE is a pure repository for proteomics experiments which does not process the submitted data in any way. PRIDE accepts submission of protein and peptide identifications as made by the submitter, together with supporting evidence, detailed information about the experimental protocol and literature references, making extensive use of ontologies and controlled vocabularies to ensure consistency of annotations. All data submitted to PRIDE is validated against current proteomics data standards. Although submitted data is not re-processed, protein and peptide identifications are periodically mapped to the latest protein database entries via the UniProt Archive [263], a warehouse of all publicly available protein sequences. This facilitates data integration within the PRIDE database as well as with external resources. Protein and peptide identifications can be accessed via a web-interface and the PRIDE BioMart [264, 265].

1.7.1.4 Human Proteinpedia

HuPA differs from the above resources in that it is not only a repository of mass spectrometry data, but also stores data from a diverse range of experimental platforms, including yeast two-hybrid screens, peptide/protein arrays, immunohistochemistry, co-immunoprecipitation and fluorescence microscopy [258]. HuPA is developed by the Johns Hopkins University, Baltimore, USA, in collaboration with the Institute of Bioinformatics, Bangalore, India. Like PRIDE, it accepts submissions of mass spectrometry experiments together with extensive ontology based meta information. MS proteomics data is integrated with data from other experimental platforms and protein annotations through the Human Protein Reference Database (HPRD) [266]. Raw data and identifications are available in different flat file and extensible markup language (XML) formats.

1.7.1.5 The ProteomExchange consortium

As each of the above repositories captures a different subset of data and currently provides it in a different form, accessing all proteomics data available in public databases still requires considerable effort. The ProteomExchange consortium has been established with the aim of setting up a regular data exchange between major proteomics repositories, including PRIDE, PeptideAtlas and gpmDB [267].

1.7.2 Protein sequence databases

1.7.2.1 UniProtKB

Protein sequence information is generated in a highly redundant fashion, resulting in sequence data corresponding to a single protein coming from many different sources such as genome projects, full length cDNA, or less frequently, protein sequencing. UniProt [268], a resource unifying the Swiss-Prot [269], TrEMBL [269] and Protein Information Resource (PIR) [270] databases, was created with the aim of providing a central database where sequences relating to the same protein are merged into a unique entry describing all unique protein products of an individual gene. The UniProt Knowledgebase (UniProtKB) provides extensive curated information on proteins, including functional annotation, classification and cross references to external databases. Increasing use of controlled vocabularies and ontologies improves data integration and programmatic data access.

UniProtKB/Swiss-Prot is supplemented by UniProtKB/TrEMBL which contains protein sequences resulting from translation of nucleotide sequences in the European Molecular Biology Laboratory (EMBL) nucleotide sequence database [271] which are then electronically annotated based on similarity to sequences in UniProtKB/Swiss-Prot.

In 2001 the UniProt consortium launched the Human Proteome Initiative (HPI) [272], a major project to annotate all known human sequences according to the Swiss-Prot quality standards. A partial success in reaching this aim was recently achieved with the availability of a manually curated representation of all currently known human protein-coding genes in Swiss-Prot.

1.7.2.2 Ensembl

Ensembl is one of three resources (Ensembl [10], University of California Santa Cruz (UCSC) GenomeBrowser [273], NCBI MapView [274]) that integrate a diverse range of information about genes, transcripts, proteins and functional DNA elements on the annotated genome sequence. Gene models in Ensembl are predicted by the Ensembl annotation pipeline [9] based on mRNA and protein sequence data from the EMBL nucleotide sequence database, UniProtKB and RefSeq [275] . Transcripts derived from the gene models are translated into protein sequences. It should be noted here that the protein sets provided by Ensembl are partially redundant on sequence level as translations of transcripts with different untranslated regions (UTR) but identical coding sequences (CDS) are propagated as separate database entries.

1.7.2.3 International Protein Index

The International Protein Index (IPI) [276] is a database that has been created to offer a complete non-redundant representation of protein sequences in the public domain. IPI entries are derived by clustering entries from primary sequence databases based on sequence similarity, as well as information from the source database indicating that they represent the same biological entity. Master sequences are chosen for each cluster according to a defined hierarchy of the source databases [276]. Initially developed for the primary analysis of the human genome sequence, IPI is now a prime database for MS based identification of proteins. A tradeoff between sequence databases of high quality but limited coverage like Swiss-Prot and more comprehensive but also more redundant databases like NCBInr, IPI offers a maximally complete but minimally redundant set of protein sequences.

1.8 Ontologies

In the context of databases, the concept of ontologies should be briefly introduced as they are of increasing importance in the organisation and processing of biological information, including functional annotation of genes and proteins. In this work ontologies are used to analyse the functional composition of protein sets.

In analogy to the philosophical concept, ontologies in the life sciences are applied to the categorisation of objects and the description of the relationships between them. Ontologies are structured and well defined controlled vocabularies allowing for a unified and consistent annotation of objects. Ontologies are formally structured as directed, acyclic graphs (DAG). In a DAG information flows only in one direction without any circular relationships between the nodes of the graph. The nodes of the graph are terms describing an object or function and the edges connecting the nodes define the relationships between the terms. The nodes of the graph are organised hierarchically such that terms describing more general concepts are *parent* nodes of more specific *child* terms.

The most important ontology to unify functional annotation of genes and gene products is the Gene Ontology (GO) [277]. It consists of three orthogonal ontologies, sometimes also referred to as categories, which permit complementary description of gene function in terms of their *molecular function*, the *biological process* they act in and the *cellular component* they are part of. The hierarchical structure of the ontology is exemplified in figure 1.5, showing the graph of the parent-child relationships for the term *translation* which is part of the *molecular function* ontology. It also illustrates the complexity of the ontology structure. Note the directionality of the graph and the absence of circular relationships.

When performing statistical tests on individual ontology terms statistical dependencies between terms introduces by the graph structure have to be taken into account and corrected for [278].

Two other ontologies of importance in the context of this work are the

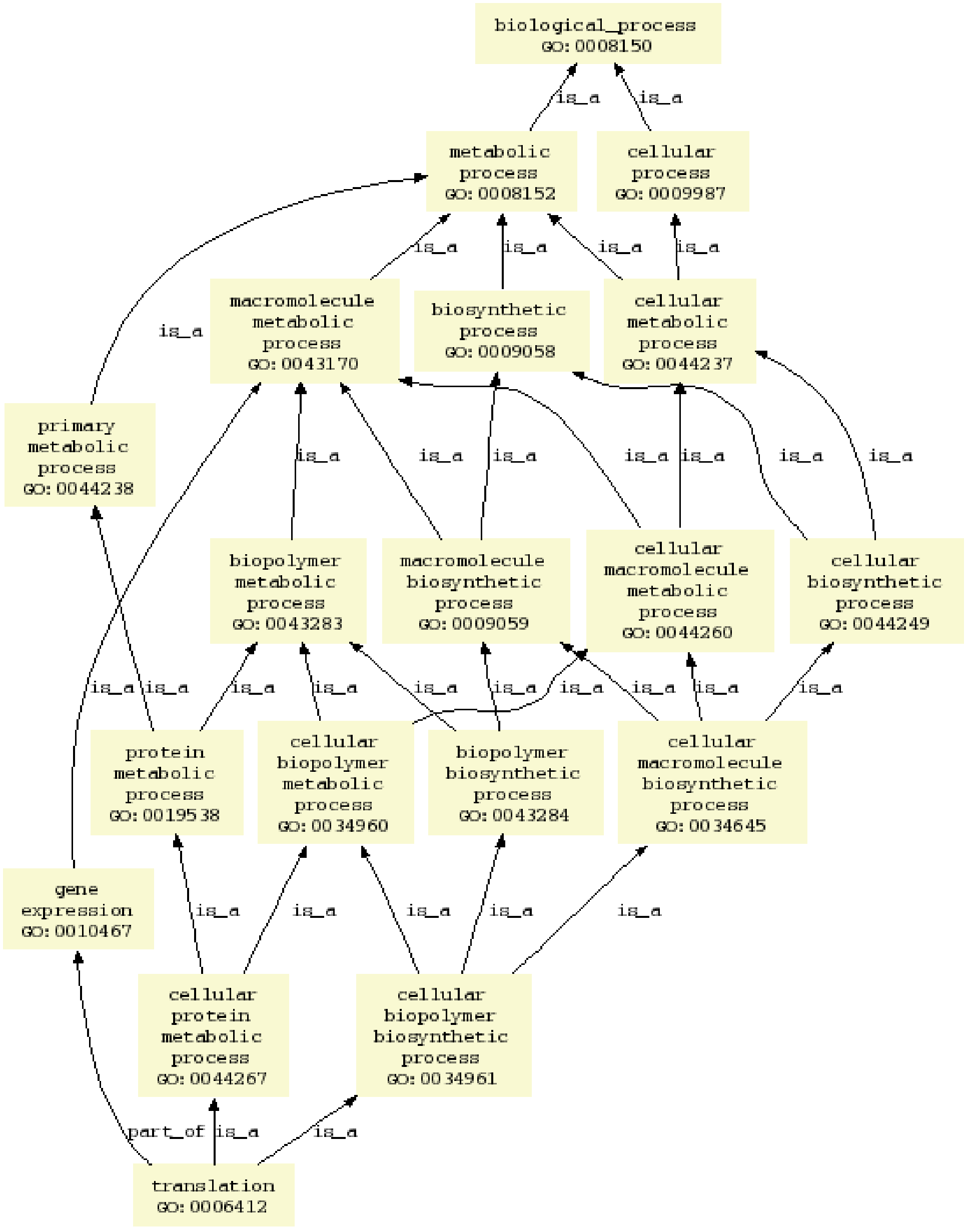

biological_process GO: 0008150
is_a
is_a
metabolic process GO: 0008152
cellular process GO: 0009987
is_a
is_a
is_a
is_a
is_a
macromolecule metabolic process GO: 0043170
biosynthetic process GO: 0009058
cellular metabolic process GO: 0044237
primary metabolic process GO: 0044238
is_a
is_a
is_a
is_a
is_a
is_a
is_a
biopolymer metabolic process GO: 0043283
macromolecule biosynthetic process GO: 0009059
cellular macromolecule metabolic process GO: 0044260
cellular biosynthetic process GO: 0044249
is_a
is_a
is_a
is_a
is_a
is_a
is_a
is_a
is_a
is_a
protein metabolic process GO: 0019538
cellular biopolymer metabolic process GO: 0034960
biopolymer biosynthetic process GO: 0043284
cellular macromolecule biosynthetic process GO: 0034645
gene expression GO: 0010467
is_a
is_a
is_a
is_a
is_a
cellular protein metabolic process GO: 0044267
cellular biopolymer biosynthetic process GO: 0034961
part_of
is_a
is_a
translation GO: 0006412

Figure 1.5: Example of the parent-child relationships of a GO term. Shown are the paths through the GO graph from the term *translation* to the root term of the *molecular function* ontology. Note the directionality and absence of circular relationships. The graph was generated by the Ontology Lookup Service [279].

eVOC ontology [280] and InterPro [281]. eVOC is used to unify the annotation of gene expression data. Ensembl transcripts are linked transitively with eVOC terms by mapping ESTs and Affymetrix probesets of eVOC annotated expression datasets to the genome. Expression information from EST libraries deposited in the dbEST database [282] and mRNA expression profiling data from the Genomics Institute of the Novartis Research Foundation (GNF) Atlas of Gene Expression [283] is used to annotate Ensembl transcripts.

InterPro describes the relationship between protein families and domains. Strictly speaking the InterPro hierarchy is a controlled vocabulary rather than an ontology as it does not have a DAG structure and contains cyclical relationships.

1.9 Dissertation outline

After this introduction to concepts, technologies and methods important in the context of the large-scale study of protein function, the following three chapters are concerned with a comprehensive evaluation of experimental and *in silico* data that aims to critically assess the performance and scope of current mass spectrometry based annotation of the human proteome. In each chapter a separate introduction to the background of the respective topic will be given.

Chapter two describes the results of an integrative bioinformatics analysis of protein expression data, generated in a large-scale collaborative study of the human brain proteome, to examine resolution, functional composition and coverage of the studied proteome.

Chapter three assesses, to what extent the findings from the analysis of the brain proteome study reflect global trends in mass spectrometry based proteomics on the basis of a comprehensive survey of proteomics data in the public domain in the context of the genome sequence.

Following on from the results obtained from the analysis of experimental data, the fourth chapter will outline and evaluate a strategy to improve the selectivity and sensitivity of targeted proteomics approaches through diversification of peptide populations by combinatorial proteolysis.

The dissertation concludes with a brief summary and outlook section.

Chapter 2

Anatomy of a collaborative large-scale proteomics study: functional analysis of proteins identified in the HUPO Brain Proteome Project pilot phase

2.1 Introduction

Availability of the human genome sequence [6] paved the way for the application of high-throughput methods to the study of the human proteome. Given the complexity of the proteome, however, the proteomics community could only take full advantage of the opportunities ahead by harnessing the power of collaborative research. With the foundation of the Human Proteome Organisation (HUPO) [284, 285] in February 2001, the proteomics community formally acknowledged the magnitude of the challenges involved in studying the human proteome. HUPO's mission statement is "To define and promote proteomics through international cooperation and collaborations by fostering the development of new technologies, techniques and training to better understand human disease" [286].

The HUPO Brain Proteome Project (BPP) is one of the currently seven initiatives dedicated to the study of cells, tissues and bodily fluids [245, 284, 244, 246, 248, 247, 244]. The aims of the BPP are to define the protein complement of the human brain in health and disease and to detect qualitative and quantitative changes associated with physiological and pathological states with a focus on neurodegenerative diseases and ageing. The long term aim is to reveal potential biomarkers for brain diseases.

The brain is the most highly developed organ of the human body characterised by its diversity of cell types and complex structure. It is thought that up to 50% of all protein-coding genes are expressed in the brain. Given current estimates of the total number of protein-coding genes in the human genome that is still only around 11000 genes. Phenotypic complexity from this relatively small gene pool is achieved by post-transcriptional and post-translational mechanisms: alternative splicing, covalent modification and proteolytic processing result in an extensive diversification at protein level and modulate protein activity, as well as the intricate network of protein interactions that adds an additional layer of complexity. Thus, findings from genomic and transcriptomic studies - that have already contributed important information to the understanding of this highly complex organ [287, 288] - have to be complemented by proteomics investigations.

Studying the composition of the brain proteome will also contribute to the understanding of the evolution of this complex organ. It has been suggested that variations at the protein level are major determinants of differences between species with very similar genome sequences, and that these variations play a particularly critical role in the evolution of the brain [289].

Finally, proteomics can help to identify proteins involved in neurodegenerative changes that cause common brain-related diseases [290]. The study of brain proteins has therefore become an important tool in medical and pharmaceutical research, and has contributed to the elucidation of the molecular mechanisms underlying neurodegenerative diseases such as Alzheimer's, Parkinson's, and multiple sclerosis [287, 288, 291, 292, 293, 294, 295].

Apart from technological challenges common to all current proteomics analyses the structural and compositional complexity of the brain poses a significant additional analytical challenge to brain proteomics. The complex organisation makes it difficult to prepare the often small substructures under study. Potential changes caused by a pathological condition could thus be overshadowed by existing differences between target and proximate areas [290].

Furthermore, human brain tissue *per se* is difficult to obtain. Several approaches therefore focus on the analysis of brain derived proteins in bodily fluids like cerebrospinal fluid (CSF) [236, 296]. CSF is widely considered the most suitable 'proximal' fluid as it is in direct contact with the central nervous system (CNS) and thus a likely candidate for the identification of disease biomarkers that reflect neuropathological changes [297, 298].

The HUPO BPP study analysed the protein content of human temporal lobe tissue. The advantage of temporal lobe tissue being that it can be obtained from living, albeit not healthy, individuals by excisional biopsy from patients undergoing temporal lobectomy, a neurosurgical intervention to treat epilepsy [299].

The temporal lobes are part of the cerebrum and located at the sides of the brain. They are involved in emotional responses, linguistic and visual semantics, aural processing and memory [300]. Temporal lobe tissue is an important subject of study for epilepsy research, as seizures in mesial temporal lobe epilepsy (MTLE) - one of two main types of epilepsy - arise from the hippocampus which is part of the temporal lobes [301]. Several proteomics studies have analysed temporal lobe tissue to investigate the causes of epilepsy [302]. Proteomic analysis of temporal lobe tissue is also relevant to the understanding of other neurological diseases such as Alzheimer's disease [303, 304, 305] and schizophrenia [306].

The HUPO BPP was launched in April 2003 [307] and is the first attempt to systematically study the brain proteome on a large scale. A pilot study to evaluate different fractionation and mass spectrometry (MS) technology platforms and to establish a central informatics infrastructure to collate, integrate, analyse and disseminate experimental data was agreed in September 2003 [308]. Nine groups participated in the study, which concluded in 2006 [132].

Participants analysed protein samples from specimens of human temporal lobe tissue obtained by autopsy and biopsy [309, 310, 311]. Biopsy specimens originated from hippocampal tissue obtained from a patient who underwent surgical treatment of MTLE. Autopsy samples were obtained from the same area from an individual with no signs of neurodegeneration. Tissue specimens from autopsies have the advantage of being "normal", as they originate from tissue that did not undergo pathological changes. A drawback of autopsy material is, however, that many proteins will have already been degraded if the autopsy was performed more than a couple of hours after death. Therefore, one objective of the HUPO BPP study was a comparative assessment of the suitability of the two different specimen types for the study of the brain proteome. Mouse brain samples taken from different developmental stages were also analysed [312, 313, 314, 315, 316] but the corresponding findings are not the subject of the analysis presented here.

All mass spectra generated in the pilot study were collated in a central data collection centre (DCC) and subjected to an automatic reprocessing pipeline to obtain consistent results across the heterogeneous datasets submitted. Data reprocessing as well as peptide and protein identification was conducted by collaborators at the Centre for Medical Proteomics in Bochum and is described in [317]. In brief, mass spectra were searched against a composite database consisting of International Protein Index (IPI) protein sequences and decoy sequences obtained by shuffling of the amino acid sequences of the original IPI entries. The false positive rate of protein identification could then be determined based on the identification rate of decoy sequences. To obtaine a set of high confidence protein identifications results from the three different search engines SEQUEST, MASCOT and ProteinSolver were combined using the custom algorithm ProteinExtractor [317].

This chapter discusses the results of an integrative bioinformatics analysis downstream of the peptide and protein identification process which was carried out with the aim to put the identified proteins into a broader biological context. To this end protein identification were integrated with existing knowledge derived from different, mainly protein-centric data sources.

First, an overview of the protein sets identified in the participating laboratories by different technologies is provided. Then the dataset is illuminated from a functional perspective by analysing it in the context of genomic and transcriptomic information, categorising protein identifications by subcelluar localisation, molecular function and biological process, grouping them into protein families and mapping them to biological pathways. The detection of proteins pivotal to brain function is assessed based on a detailed analysis of peptide evidence. Since splice isoforms can play an important role in neurodegenerative diseases such as Alzheimer's [318, 319, 320] peptide evidence is also analysed with regard to splice isoform resolution for brain specific isoforms. Finally, brain specific peptide evidence from the HUPO BPP study is compared to that from a large scale proteomics study in CSF [236] to investigate the detection of brain derived proteins in CSF.

2.2 Materials & methods

2.2.1 Protein and peptide identifications

HUPO BPP protein and peptide identifications resulting from centralised reprocessing of the collected mass spectra [317] along with sample metadata were obtained from the Protein Identification Database (PRIDE) in form of nineteen sets of IPI entries identified in the respective experiments together with the peptide identifications used for protein inference. Details of the datasets are shown in table 2.1.

Table 2.1: HUPO BPP experiments in human temporal lobe tissue. Details of the 19 experiments that identified proteins in human temporal lobe specimens. Given for each experiment are an experiment ID (exp. ID), an ID for the laboratory that generated the mass spectra (lab ID), the specimen type the protein sample was obtained from, the separation technique, the MS device used for peptide identification, the MS technology of the device (MS tech.), the number of proteins identified (ident. freq.) and the size of the non-redundant set of proteins identified (ident. freq. nr).

exp. ID	lab ID	specimen type	separation technique	MS device	MS tech.	ident. freq.	ident. freq. nr
07	02	biopsy	2D-PAGE	ABI 4700	TOF	31	7
08	02	autopsy vs biopsy	2D-PAGE	ABI 4700	TOF	1	1
14	10	autopsy vs biopsy	2D-LC	Finnigan LTQ	IT IT	234	234
16	10	autopsy	2D-LC	Finnigan LTQ	IT	406	406
17	10	autopsy	2D-LC	Finnigan LTQ	IT	481	481
18	10	autopsy	2D-PAGE	ABI 4700	TOF	72	63
15	10	autopsy vs biopsy	2D-PAGE	ABI 4700	TOF IT	5	3
29	12	biopsy	1D-PAGE	Finnigan LCQ Classic	IT	45	45
26	12	biopsy	2D-PAGE	Finnigan LCQ Classic	IT	47	29
28	12	biopsy	2D-PAGE	Finnigan LCQ Classic	IT	30	28
27	12	biopsy	2D-LC	Finnigan LCQ Classic	IT	169	169
33	12	autopsy	1D-PAGE	Finnigan LCQ Classic	IT	55	55
30	12	autopsy	2D-PAGE	Finnigan LCQ Classic	IT	47	47
32	12	autopsy	2D-PAGE	Finnigan LCQ Classic	IT	53	46
31	12	autopsy	2D-LC	Finnigan LCQ Classic	IT	118	118
34	12	autopsy	2D-LC	Finnigan LCQ Classic	IT	137	137
35	13	biopsy	1D-PAGE	Finnigan LTQ FT	IT FT	1235	1235
36	13	biopsy	1D-PAGE	Finnigan LTQ FT	IT FT	480	480
37	14	autopsy	2D-PAGE	Bruker Ultraflex	TOF	322	202

Protein and peptide identifications made by Pan *et al.* in CSF [236] were also obtained from PRIDE. Details of the experiments are shown in table 2.2.

Table 2.2: Experiments in human CSF. Details of the experiments that identified proteins in human cerebrospinal fluid. For each experiment an experiment ID (exp. ID), the separation technique, the MS device used for peptide identification, the MS technology of the device (MS tech.), the number of proteins identified (ident. freq.) and the size of the non-redundant set of proteins identified (ident. freq. nr) are given.

exp. ID	separation technique	MS device	MS tech.	ident. freq.	ident. freq. nr
212	2D-LC	Finnigan LCQ	IT	357	357
217	2D-LC	Finnigan LTQ	IT	535	535

2.2.2 Database release mapping

2.2.2.1 Protein and peptide mapping

IPI entries were mapped from IPI release 3.05 to releases 3.14 and 3.29 using the IPI history file available from the IPI database file transfer protocol (FTP) server (ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/ipi.HUMAN.history.gz). Peptide identifications were mapped to sequence entries of IPI release 3.29 by identical string matching. A maximal explanatory set of proteins was obtained by retaining all IPI entries with at least one peptide match against their sequence. No enzyme constraints were enforced. The set was mapped to Ensembl protein and gene entries using cross reference information obtained from the IPI database FTP server (ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/ipi.HUMAN.xrefs.gz).

2.2.3 Protein annotations

2.2.3.1 Sorting signals and transmembrane domains

Signal sequences, transit sequences and transmembrane domain annotations were obtained from the UniProt Knowledgebase (UniProtKB) [321] and Ensembl [322]. Sequences without information in either of the databases were subjected to the respective prediction algorithms: presence of signal and transit sequences was predicted with SignalP 3.0 [323] and TargetP 1.1 [204] respectively. TMHMM 2.0 [324] was used to predict transmembrane domains.

2.2.3.2 Protein domains and families

Assignment of proteins to protein families was based on InterPro [325] domain signatures identified in the sequence by InterProScan [326]. InterPro annotations for IPI entries were obtained from the IPI FTP server (ftp.ebi.ac.uk/pub/databases/IPI/current/ipi.HUMAN.IPC.gz).

2.2.3.3 Splice isoforms

IPI entries that were associated with the same Ensembl gene through their Ensembl cross reference, were considered splice isoforms of the respective gene. Here, splice isoforms are defined as proteins encoded by the same gene that differed in their protein sequence as a result of alternative splicing. Proteins which were translations of different transcripts which differed only in their untranslated region (UTR) and thus encode the same protein product were considered a single splice isoform. The number of splice isoforms on transcript level could therefore differ from the number of splice isoforms on protein level.

Based on this definition the splice rate of a gene was given by the number of distinct translations associated with an Ensembl gene through cross references to the respective Ensembl protein entries.

2.2.3.4 Tissue expression

Analysis of tissue expression was based on the annotation of Ensembl transcripts with terms from the eVOC ontology [280], a controlled vocabulary to describe gene expression data. eVOC annotations of Ensembl transcripts were obtained from the Ensembl BioMart [327] MySQL database (ensembldb.ensembl.org/ensembl_mart_36).

2.2.3.5 Functional categorisation

Functional categorisation of proteins was based on association of IPI entries with Gene Ontology (GO) terms [277]. Assignments of IPI entries to GO terms were obtained from the Gene Ontology Annotation (GOA) database [328] release 38. The gene association file containing the assignments was downloaded from the GOA FTP server (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa).

2.2.3.6 Biological pathways

The distribution of identified proteins across biological pathways was analysed using information available from the Reactome database [329]. The content of Reactome version 16 was downloaded as a MySQL dump and imported into a local MySQL database which was subsequently queried for the relevant information using UniProtKB cross references of identified IPI entries. A graphical representation of the distribution of identifications across Reactome pathways was generated using the SkyPainter tool [330] at http://www.reactome.org/cgi-bin/skypainter2.

2.2.3.7 Disease association

Associations of IPI entries with entries in the Online Mendelian Inheritance in Man (OMIM) database [331] were obtained for the subset of IPI entries with cross references to UniProtKB/Swiss-Prot. OMIM cross references were extracted from the respective UniProtKB/Swiss-Prot entries.

2.2.4 Experiment clustering

A binary $p \times e$ incidence matrix I of proteins and experiments was constructed with a row for each protein p and a column for each experiment e. The entry $I_{p,e}$ had a value of 1 if protein p was identified in experiment e and a value of 0 otherwise.

Subsequently the pairwise euclidean distance between the experiment vectors (matrix columns) was calculated and used as the distance measure to cluster the experiments by hierarchical clustering.

All calculations were made using the statistical programming package *R* [332]. The euclidean distance was calculated with the `dist` function using the default parameters. The resulting distance matrix was used to calculate the clustering of the experiment vectors with the `hclust` function using the default paramaters.

2.2.5 Association with transcript expression profiles

Transcript expression levels of genes encoding proteins identified in the HUPO BPP study were obtained from the Genomics Institute of the Novartis Research Foundation (GNF) SymAtlas [229]. gcRMA normalised expression data [333] was downloaded from the SymAtlas website (http://wombat.gnf.org/downloads/gnf1h,gcrma.zip). Expression profiles measured on a HG-U133A Affymetrix gene chip were linked with transcript entries of Ensembl release 36 based on mappings of Affymetrix probesets to Ensembl transcripts obtained from Ensembl BioMart version 36. GNF custom probesets were mapped to Ensembl transcripts using the GNF chip annotation tables obtained from http://wombat.gnf.org/downloads/gnf1h-anntable.zip. IPI entries were associated with mRNA expression levels transitively through Ensembl cross references obtained from the IPI FTP server.

2.2.6 Analysis of splice isoform identification

To analyse the specificity of splice isoform identification, the identified IPI entries were grouped by the Ensembl genes they were associated with. For each set of IPI entries the peptides used for protein identification were mapped by exact string matching to all IPI entries - both identified as well as unidentified - associated with the respective Ensembl gene.

2.2.7 Gene Ontology analysis

For each GO term annotation the reflexive transitive closure was calculated. That is IPI entries associated with a specific term were recursively mapped to all (more general) ancestor terms in the hierarchy. GO term over- and under-represention was then assessed by calculating the hypergeometric distribution function $F_{hyper}(x_t)$ for each observed GO term frequency x_t: Let N be the number of IPI entries annotated with at least one GO term, n the number of proteins that are annotated with a specific GO term t, M the fraction of N that has been identified and x_t the fraction of M that is associated with GO term t. The probability P_U of observing x_t or fewer proteins associated with GO term t in a sample of size M is then calculated according to equation 2.1.

$$P_U = F_{hyper}(x_t) = P(X \leq x_t) = \sum_{i=0}^{x_t} \frac{\begin{pmatrix} M \\ x \end{pmatrix} \begin{pmatrix} N-M \\ n-i \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \tag{2.1}$$

The probability P_O of observing more than x_t proteins associated with GO term t is then calculated according to equation 2.2.

$$P_O = 1 - P_U \tag{2.2}$$

P values were calculated using the statistical analysis suite R [332]. Significance of GO term over-representation was assessed by P_O and significance of GO term under-representation by P_U.

The significance threshold was chosen by controlling the false discovery rate (FDR) at 0.01% using the linear step-up procedure by Benjamini and Hochberg [334]. This procedure makes use of the ordered p-values $P_1 \leq ... \leq P_m$: denote the corresponding null hypotheses $H_1, ..., H_m$. For a desired FDR level q, the ordered p-value P_i is compared to the critical value $q \cdot i/m$. Let $k = max\{i : P_i \leq q \cdot i/m\}$. Then reject $H_1, ..., H_k$ if such a k exists [335].

Only GO terms assigned to ten or more proteins were considered in the analysis. It should be noted that due to multiple testing issues and dependencies between GO terms the obtained *P* values have to be interpreted with caution.

2.2.8 Identification of genes putatively pivotal to CNS function

A list of genes putatively important for CNS function was compiled based on three sources of information:

1. biomedical literature
2. curated annotation
3. mRNA expression profiles

2.2.8.1 Biomedical literature

The biomedical literature was queried via PubMed (http://www.pubmed.org) using the search expression:

```
(brain[TI] AND specific[TI] AND splice[TI])  OR
(brain[TI] AND specific[TI] AND isoform[TI]) OR
(alternative splicing[MeSH] AND brain[MeSH] AND humans[MeSH])
```

Gene names were manually extracted from the title and abstract of the retrieved publication. The gene names were subsequently mapped to Ensembl gene identifiers by querying the Ensembl database with the obtained gene names. Gene names not mappable through Ensembl directly were first converted to approved Human Genome Organisation Genome Nomenclature Consortium (HGNC) [336] gene symbol by searching the HGNC database (http://www.genenames.org) [337]. The obtained HGNC symbols were then used to query Ensembl.

2.2.8.2 Curated annotation

The UniProtKB/Swiss-Prot database was queried with a keyword search for 'brain'. The retrieved entries were then filtered manually based on the comment line topic 'tissue specificity' [338] for proteins annotated as brain specific or having a brain specific splice isoform. The obtained UniProtKB/Swiss-Prot accession numbers were mapped to Ensembl gene identifiers using Ensembl BioMart.

2.2.8.3 Transcript expression

Genes expressed preferentially in CNS tissues were identified using the GNF SymAtlas datasets [229]. gcRMA normalised expression data [333] was downloaded from the SymAtlas website (http://wombat.gnf.org/downloads/gnf1h,gcrma.zip) and screened for expression profiles showing maximum expression in adult CNS related tissue and $\leq$ 10% of the maximum expression value in all other tissues. Embryonic tissues were excluded from the search. The obtained Affymetrix probeset identifiers were mapped to Ensembl gene identifiers using Ensembl BioMart. GNF custom probeset identifiers were mapped to Ensembl gene identifiers using the GNF chip annotation tables (http://wombat.gnf.org/downloads/gnf1h-anntable.zip).

2.2.9 Monte Carlo simulation to assess enrichment of genes putatively pivotal to CNS function

A set of IPI entries indentical in size to the number of IPI entries matched by peptides identified in the HUPO BPP study was randomly sampled from IPI version 3.29. The selected IPI entries were mapped to Ensembl gene entries. Subsequently, the intersection between the randomly selected gene set and the set of genes identified as described in section 2.2.8 was calculated.

2.2.10 Permutation testing to assess the performance of proteotypic peptide predictions

A set of peptides was sampled randomly from the population of peptide sequences unique across the IPI database. A combination of separation and detection method sampled from the pool of combinations used in the HUPO BPP experiments was then assigned randomly to each peptide. For each peptide the overlap between randomly assigned separation and detection method was compared to the predictions made by Mallick *et al.* [339]. Average random overlap was calculated over ten thousand permutations.

2.3 Results

2.3.1 Database release mapping

Central reprocessing of mass spectra generated by the participating laboratories in human specimens during the HUPO BPP pilot study produced nineteen sets of protein identifications (see table 2.1) containing a non-redundant set of 1832 proteins.

IPI version 3.05 was used as the sequence database for protein identifications during central reprocessing. To allow integration with the most recent annotations, identifications were first mapped to version 3.14, the current IPI version at the time of the analysis. The results of the mapping are summarised in table 2.3.

While the majority of identified IPI entries (1737 [94.8%]) remained unchanged, 95 IPI entries had changed between the two database releases. Of these, 93 were propagated to new entries which in 23 cases resulted in merging with one or more entries in the original set. Two entries had been removed from the database because the master sequence of the IPI cluster had become defunct in the source database. The mapped set of identifications contained 1806 IPI entries.

Table 2.3: Summary of database release mapping. For each mapping status the table shows the number of database entries it applies to in the source database release, and the number of resulting entries in the target database release.

mapping status	entry count IPI V3.05	entry count IPI V3.14
Unchanged	1737	1737
Propagated	93	66
Removed	2	-
Total Mapped	1830	1806

2.3.2 Overview of protein identifications in autopsy and biopsy specimens of human temporal lobe tissue

2.3.2.1 Identification by experiment, laboratory, MS approach and specimen type

Autopsy and biopsy specimens of temporal lobe tissue were analysed by five laboratories in nineteen separate experiments using different gel-based and gel-free mass spectrometry approaches. Figure 2.1 shows the identification frequency by experiment, laboratory, separation technique and specimen type.

The bar plots show that the identification frequency varies greatly on all four levels. The number of proteins identified in the different experiments ranges from a minimum of 1 to a maximum of 1299. On average 197 proteins have been identified per experiment. The geometric mean lies at 62 identifications.

The identification frequency across the participating laboratories shows a similar level of heterogeneity. While lab #02 has contributed only eight identifications, lab #13 identified 1393 proteins. Experiments by labs #10, #12 and #14 led to 710, 344 and 199 identifications respectively.

Categorisation of identifications by separation technique shows that the majority of proteins have been identified in experiments employing one-dimensional (1D)-polyacryl amide gel electrophoresis (PAGE) followed

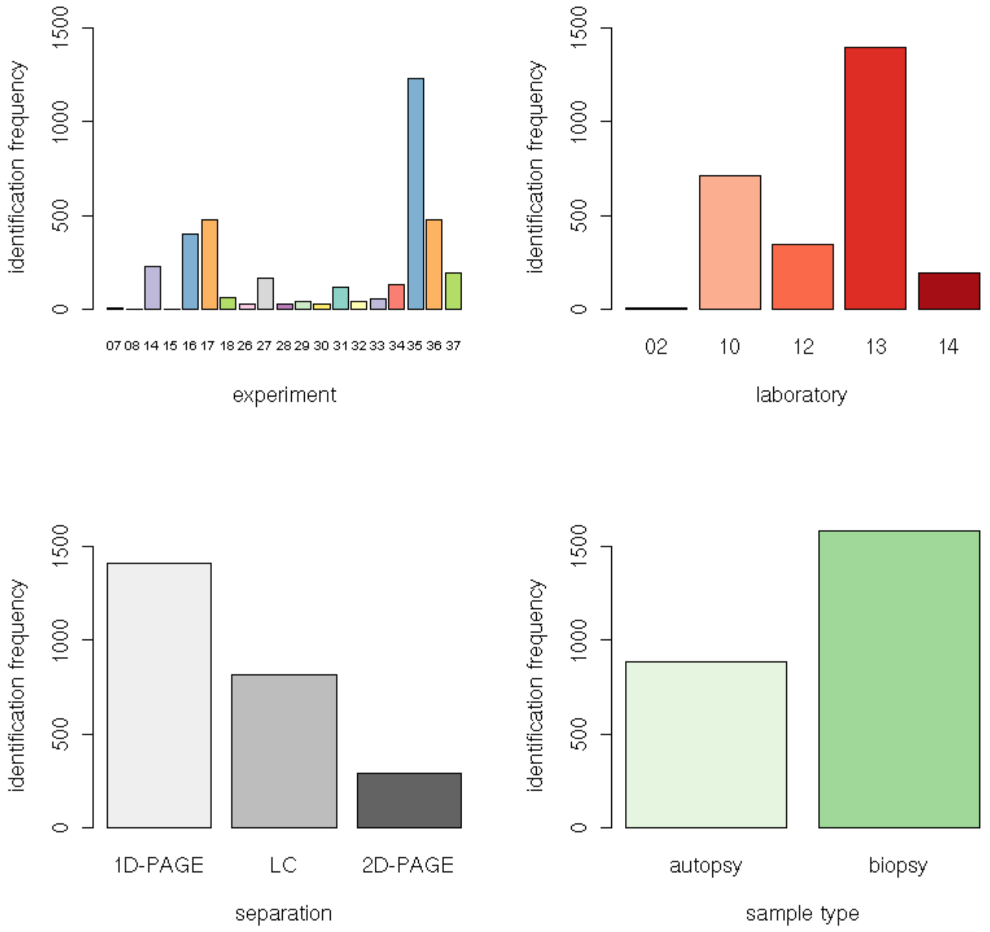


Figure 2.1: HUPO BPP protein identifications by experiment, laboratory, separation technique and specimen type. The bar plots show the number of distinct IPI entries identified.

by liquid chromatography and experiments using liquid chromatography without prior separation. The smallest number of identifications were made in two-dimensional (2D)-PAGE based experiments. A possible explanation for this is that the majority of identifications on 2D gels originate from differential analyses of protein expression in autopsy vs biopsy specimens narrowing down the set of identified proteins to those showing differential expression.

On specimen level, the number of identifications in biopsy specimens exceeds the number of identifications made in autopsy tissue by 79%.

While some labs analysed only one type of specimen using a single approach, others identified proteins in both specimen types using different methods. Table 2.4 summarises the combinations of specimen type, separation and MS technology, the laboratories that have employed them and the number of proteins identified respectively.

Table 2.4: Combinations of separation and MS technologies employed in the HUPO BPP to analyse biopsy and autopsy samples. For each combination the lab that employed the combination (lab ID), the samples analysed (sample ID) and the number of proteins identified by the respective combination are shown. Ident. freq. = the number of protein identifications reported, ident. freq. nr = the number of non-redundant proteins identified. In case of experiments that compared biopsy vs autopsy expression, samples were assigned to both the autopsy as well as biopsy categories. The respective sample identifiers and identification frequencies are underlined.

specimen type	separation technique	MS tech.	lab ID	sample ID	ident. freq.	ident. freq. nr
autopsy	1D-PAGE	IT	12	33	55	55
	2D-PAGE	IT	12	30;32	47;46	68
		TOF	2	7;8	7;1	220
			10	15;18	3;63	
			14	37	202	
	LC	IT	10	14;16;17	232;406;480	782
			12	31;34	118; 137	
biopsy	1D-PAGE	IT	12	29	45	45
		IT FT	13	35;36	1234;460	1393
	2D-PAGE	IT	12	26;28	29;28	50
		TOF	2	7;8	7;1	9
			10	15	3	
	LC	IT	10	14	232	360
			12	27	169	
autopsy vs biopsy	2D-PAGE	TOF	2	7;8	7,1	9
			10	15	3	
	LC	IT	10	14	232	232

Figure 2.2 shows the number of unique protein identifications if the samples are grouped by the combination of specimen and MS approaches. The bar plot shows that the number of identifications varies greatly across the different approaches analysing the same sample type and also between sample types analysed with the same method.

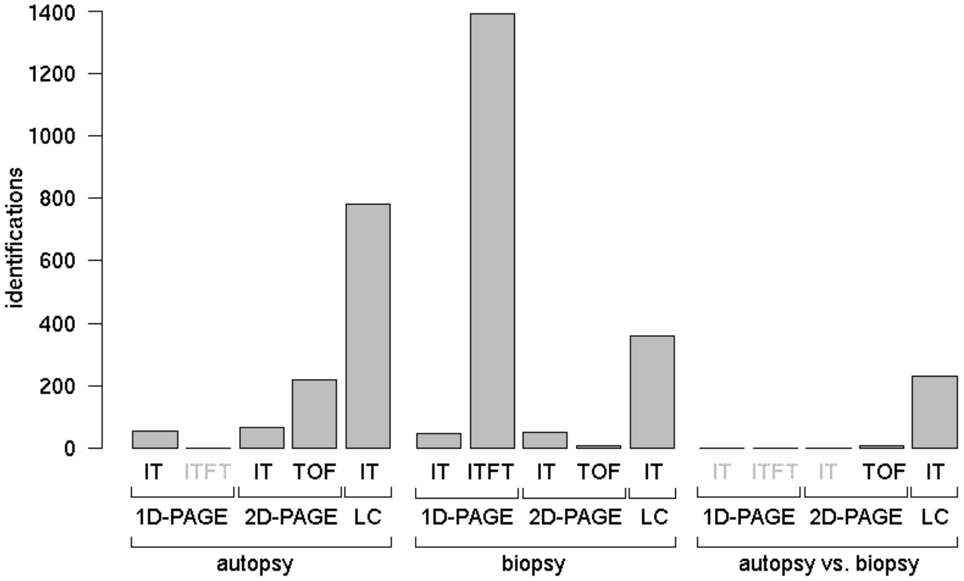


Figure 2.2: Protein identifications in biopsy and autopsy samples by the different gel-based and gel-free MS approaches employed in the HUPO BPP pilot study.

The highest number of proteins (∼1400) have been identified in biopsy tissue by a combination of 1D-PAGE and ion trap (IT)-Fourier transform (FT) MS employed by lab #13 only. Autopsy samples were not analysed by this approach. Around 800 identifications have been made in autopsy samples by liquid chromatography (LC) coupled to an IT instrument used by labs #12 and #10. The same combination yielded ∼400 identifications in biopsy samples. A further 200 identifications were made in autopsy samples separated by 2D-PAGE and analysed with time-of-flight (TOF) MS. In biopsy samples a mere nine identifications have been made using this combination of separation and MS technology.

2.3.2.2 Comparison of experiments based on protein content

To compare the protein sets identified by different MS approaches in the two specimen types an incidence matrix of protein identification versus experiment was constructed for the nineteen different protein sets (see figure 2.3).

Similarity of the protein sets identified in the respective experiments was assessed by computing the euclidian distance between the vectors defined by the sets followed by hierarchical clustering based on the calculated distances. The protein sets were ordered by the clustering result shown as a dendrogram at the top of the plot. Protein identifications were ordered by their identification frequency across experiments indicated by the colour code on the left hand side of the matrix.

Identification frequency across experiments differed greatly between proteins and ranges from one to twelve experiments. The majority of proteins (1055 [58%]) were observed only once. A small number, three proteins, were observed across twelve different experiments.

The clustering results suggest that separation technique is a strong determinant of experiment clustering. However, a comparison of the row densities of close experiments indicates that clustering is strongly influenced by the size of the different protein sets. This is visualised in figure 2.4 which shows the clustering result in context of the identification frequency.

2.3.3 Biological context of identified proteins

To put the list of 1806 proteins identified in the HUPO BPP pilot study in a biological context, the dataset was integrated with functional annotations from a range of biomedical data sources.

2.3.3.1 Annotation status

Figure 2.5 summarises the representation of proteins identified in the HUPO BPP pilot study across the different IPI source databases. Shown is

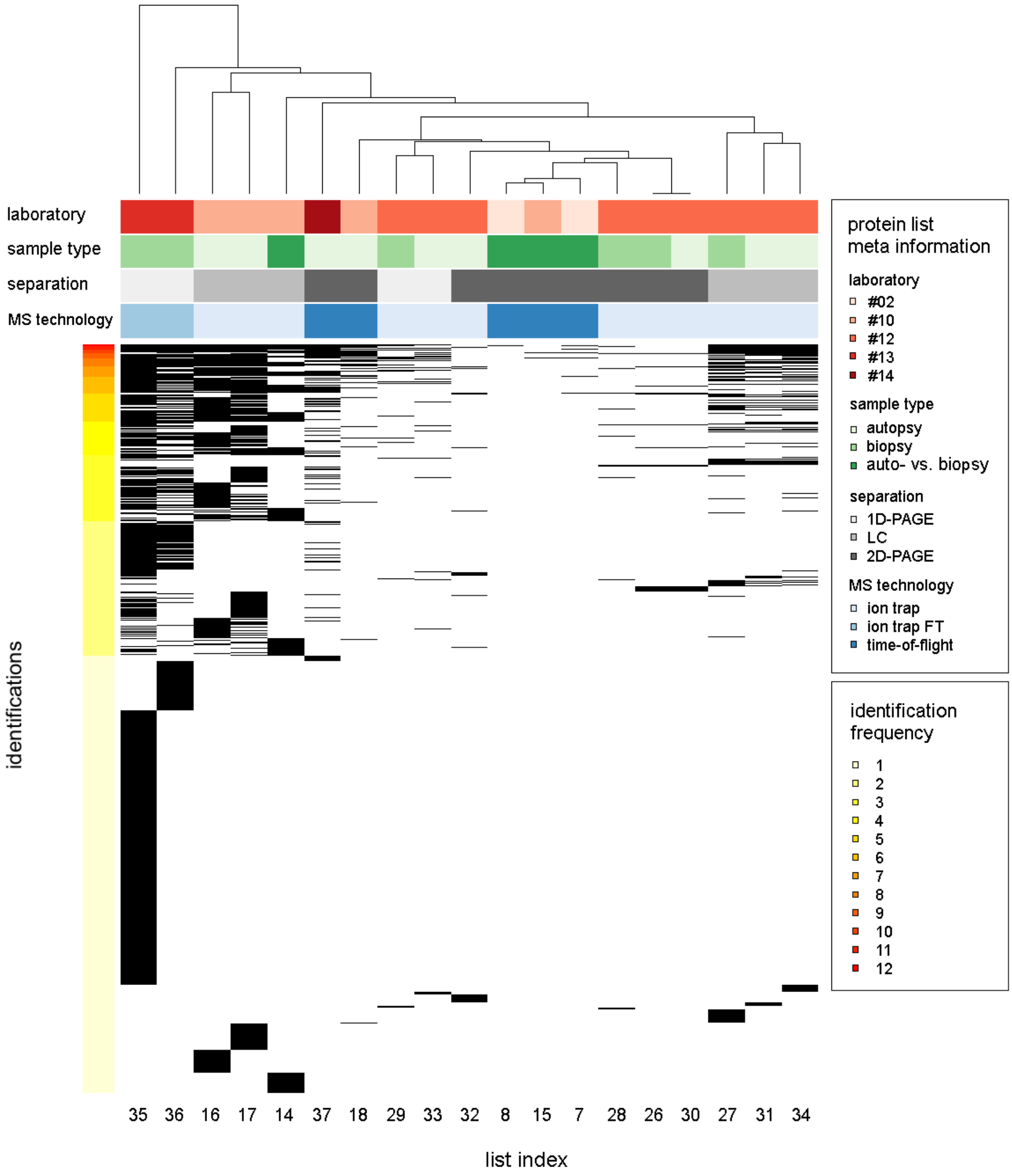


Figure 2.3: Comparison of protein identifications made in biopsy and autopsy samples by different gel-based and gel-free MS approaches. Incidence matrix of protein identifications (rows) vs experiments (columns). Horizontal bars indicate the identification of a protein in an experiment. Proteins are sorted in descending order of their identification frequency indicated by the colour on the left hand side of each row. Experiment metadata is indicated by the colour code above each row. Experiment clustering is shown by the dendrogram above the matrix. See section 2.2.4 for details of the clustering method.

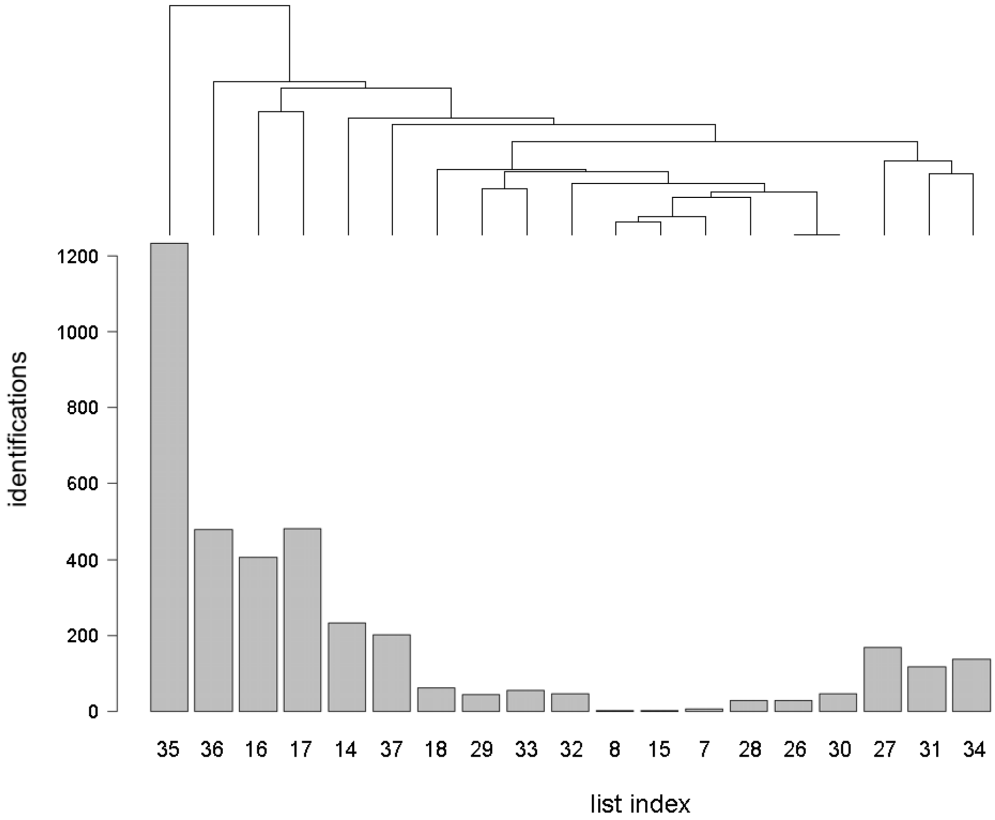


Figure 2.4: Clustering of experiments in the context of identification frequency. The incidence matrix shown in figure 2.3 is replaced by a bar plot of identification frequencies in the different experiments.

the frequency of identified IPI entries having at least one cross reference in the IPI source databases ordered hierarchically as shown on the plot. For example a protein that has a cross reference in UniProtKB/Swiss-Prot contributes to the UniProtKB/Swiss-Prot frequency and is not counted for any of the other databases, a protein that has no cross reference in UniProtKB/Swiss-Prot but in VEGA will be counted for VEGA and none of the other databases *etc.* Frequencies are plotted for all entries in the IPI database, the entries identified in the HUPO BPP study, as well as the BPP identifications grouped by separation technique.

The annotation status of a protein can to some extent be inferred from the

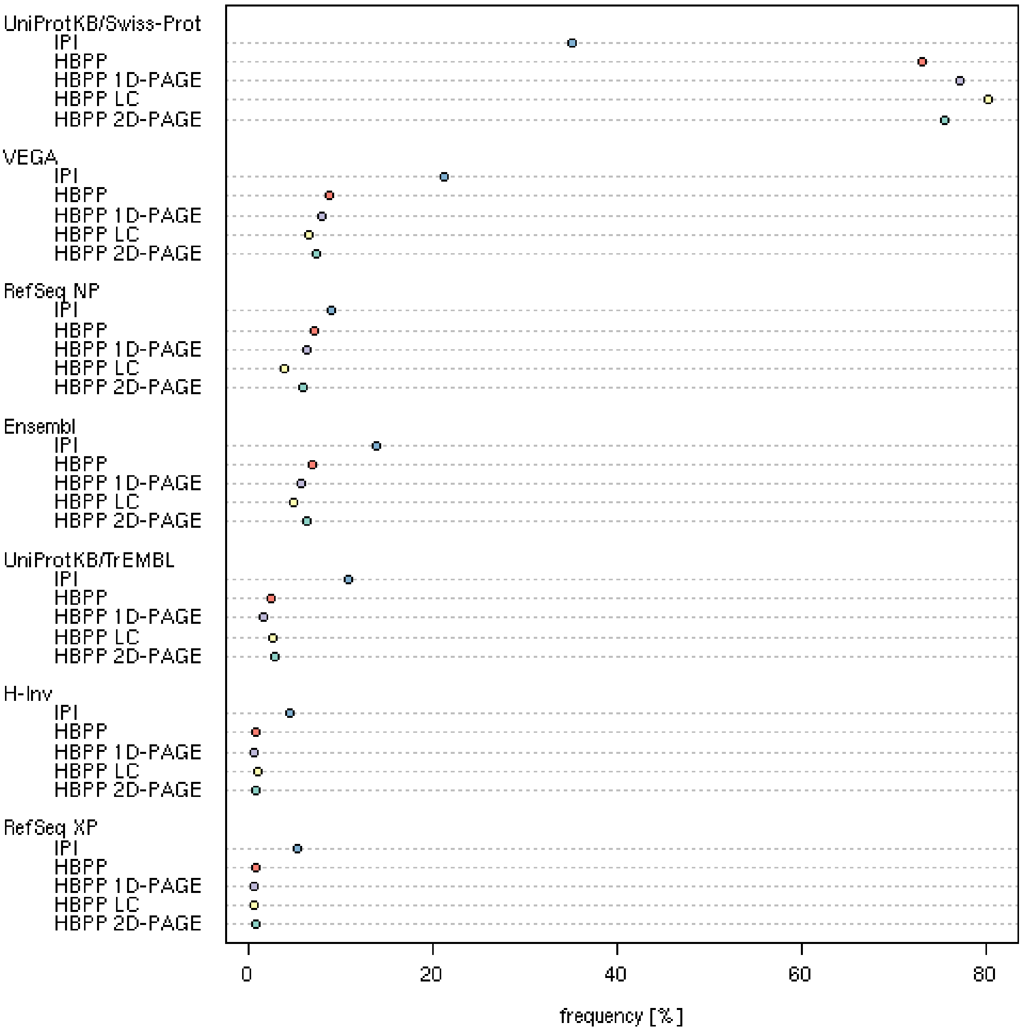


Figure 2.5: Representation of proteins identified in the HUPO BPP pilot study across IPI source databases. Shown is the frequency of identified proteins having at least one representative in the respective IPI source databases in the (top-to-bottom) order shown on the plot. See section 2.3.3.1 for a detailed explanation.

type of sequence database it is found in: proteins represented in manually curated databases like UniProtKB/Swiss-Prot tend to be relatively well known and characterised, while protein sequences resulting from unreviewed translation of complimentary DNA (cDNA) sequences or computationally predicted gene models tend to be novel and less well or not at all characterised.

The databases in figure 2.5 are loosely ordered by the level of curation and supporting evidence their entries are based on. UniProtKB/Swiss-Prot is a high quality database containing exclusively manually curated protein entries. VEGA and RefSeq NP entries are manually reviewed, while Ensembl protein entries are translated from gene models predicted computationally from various lines of evidence. UniProtKB/TrEMBL, H-Inv, and RefSeq XP entries are translations of nucleotide sequences and unreviewed gene models, respectively.

The most striking difference between the subset of IPI entries identified in the HUPO BPP study and the whole of IPI was the number of entries represented in Swiss-Prot, the manually curated part of UniProtKB. Almost 80% of identified proteins in the HUPO BPP pilot study had a corresponding entry in UniProtKB/Swiss-Prot, compared to only around 40% for the whole of IPI. On the other hand only a relatively small proportion of 4% of identified protein sequences were translations of predicted transcripts (RefSeq XP) or cDNA sequences (UniProtKB/TrEMBL, H-InvDB). In IPI overall, 20% of entries were exclusively derived from these three databases. Furthermore, the plot suggests that the database distribution is independent of the separation technique.

2.3.3.2 Genomic context

The HUPO BPP dataset was integrated with genomic information by mapping the identified IPI entries to entries in the Ensembl genome database. 1509 identifications had a cross reference to Ensembl protein entries associating them with 1420 Ensembl genes. Of these 1395 were categorised as known and 25 as hypothetical genes. The latter are gene models based on evidence from closely related species for which there is no species (human) specific evidence available. The peptides identified in the HUPO BPP study provided experimental evidence for the existence of these genes as well as their expression on protein level. Due to the absence of corresponding Ensembl protein entries 297 (16%) identifications could not be mapped to Ensembl.

In IPI the 1806 identifications are associated with 1773 IPI gene entries. The relatively large discrepancy between IPI and Ensembl gene numbers is the result of firstly, not all identified IPI entries being represented in Ensembl and secondly, the definition of 'gene' in IPI which differs from its biological meaning. Sometimes IPI entries are assigned to separate IPI gene entries even if they might originate from the same gene but merging them into one IPI gene entry would lead to contradictory IPI cross references. These contradictions are the result of inconsistencies in the protein-to-gene associations between the IPI source databases.

2.3.3.3 Identification of splice isoforms

Alternative splicing provides a powerful mechanism to establish translational complexity from a relatively small number of genes, and the importance of splicing in human genome biology has been underlined recently [340]. Splice isoforms are likely to differ in function and subcellular localisation, and often show differential tissue distribution. Isoform identification in proteomics remains challenging, however, due to the degeneracy of the identified peptide sequences [52].

Figure 2.6 shows the splice isoform frequency for genes encoding proteins identified in the pilot study compared to the whole of Ensembl. Consistent with the splice rate on genome level, the majority of Ensembl proteins identified in the HUPO BPP are the product of genes with only one distinct translation. However, when examined in more detail, there seemed to be a slight bias towards the identification of alternatively spliced genes with splice rates ranging from two to ten translations per gene in the HUPO BPP.

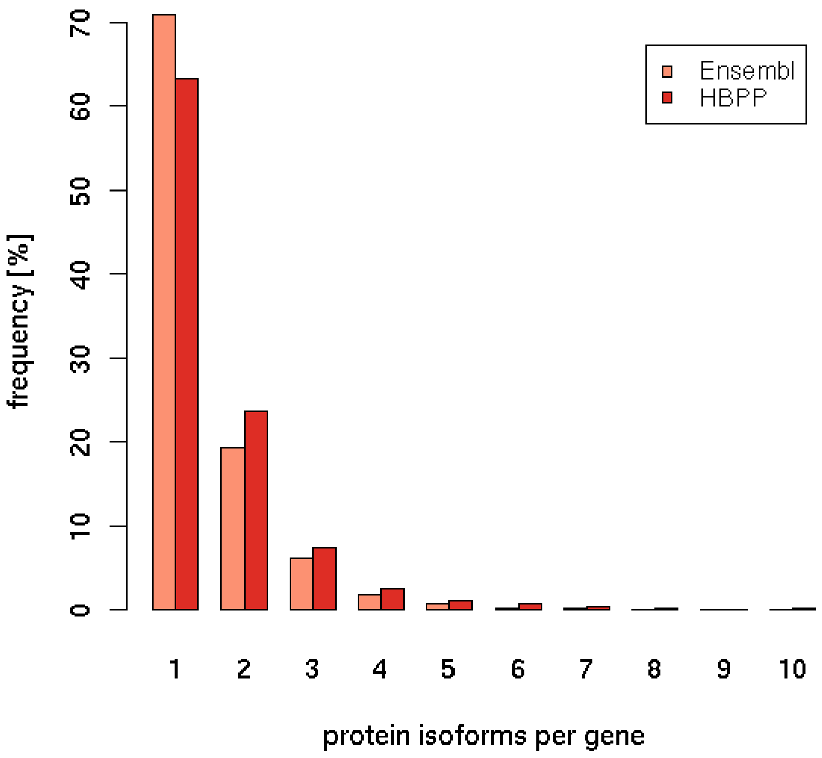


Figure 2.6: Splice rate of genes encoding proteins identified in the HUPO BPP pilot study (HBPP) in comparison to all protein-coding genes in Ensembl. The splice rate is based on the number of distinct protein sequences associated with a gene. Note that the isoform frequency on protein level differs from the splice rate on transcript level as genes might have more alternative transcripts then alternative protein products. This is the result of transcripts differing only in their UTRs and thus encoding the same protein sequence.

Figure 2.7 summarises splice isoform identification for genes encoding alternative protein products. Shown is the number of protein isoforms encoded by a gene versus the number of isoforms identified. The height of the bars is proportional to the observation frequency of each case. For example, for 54 (42 + 12) genes two of the two encoded protein isoforms were reported as identified while in 280 cases (93 + 187) only one isoform is present in the HUPO BPP dataset. In some cases all protein products encoded by a gene were identified. However, for the majority of alternatively spliced genes only a small fraction of all isoforms is reported present in the dataset.

To evaluate the specificity of splice isoform identification the distribution of peptide matches across all translations of an identified gene was determined. Peptides could then be categorised into three classes:

1. peptides matching all translations of a gene
2. peptides matching only a subset of translations
3. peptides matching exactly one translation

The latter class unambiguously identifies a particular splice isoform and is thus highly informative. The proportion of splice isoforms identified by one or more unambiguous peptide hits is shown in figure 2.7 by the area of the bar shaded in dark blue. Of the 519 protein products of alternatively spliced Ensembl genes, 219 (52%) were identified unambiguously by isoform specific peptide evidence.

2.3.3.4 Evidence for hypothetical genes and transcripts

Of the 1509 identifications that could be associated with an Ensembl gene, 1440 were translations of known, and 69 of hypothetical, Ensembl transcripts. Amongst the unambiguously identified sequences were 30 translations of hypothetical Ensembl transcripts, of which 22 are encoded by hypothetical genes (see table 2.5). Eight were hypothetical transcripts of known genes (see table 2.6).

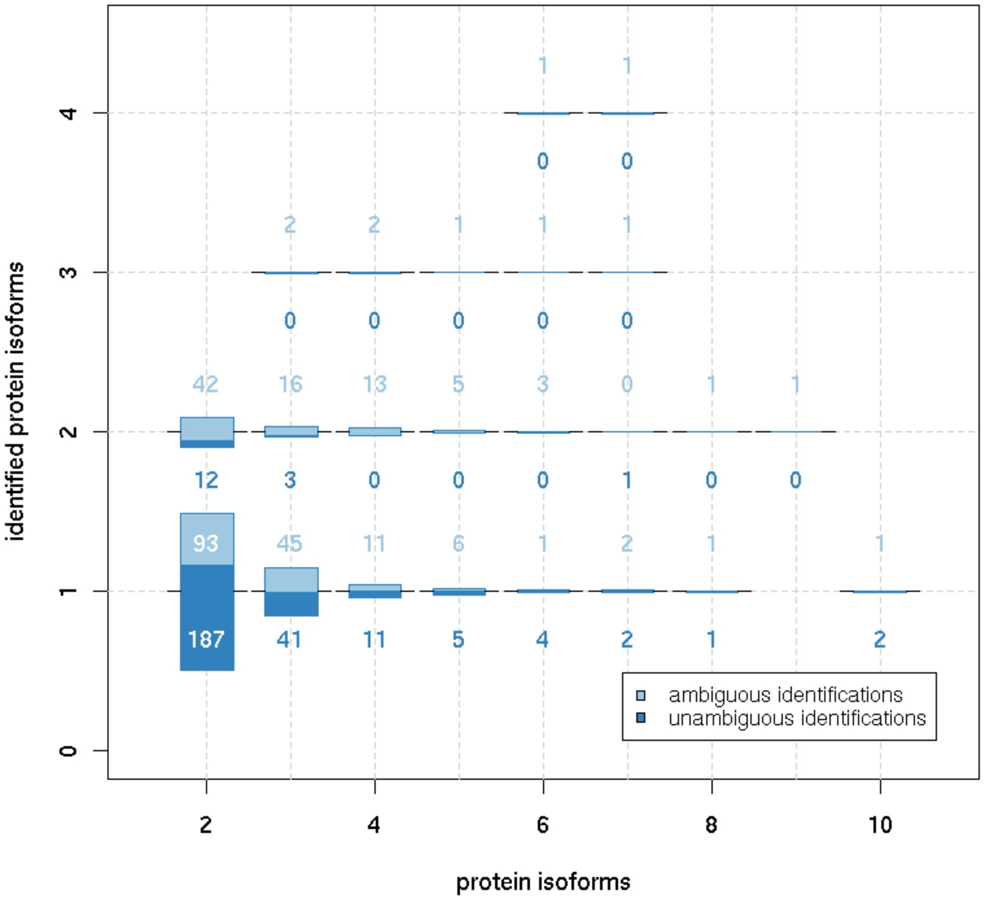


Figure 2.7: Identification of splice isoforms. Shown is the number of protein isoforms encoded by a gene as derived from the Ensembl database (x-axis) versus the number of identified isoforms (y-axis). The height of the bars is proportional to the observation frequency of each case. The proportion of isoforms identified by unambiguous peptide hits is shown in dark blue. The proportion of isoforms identified only by ambiguous peptides is shown in light blue. The corresponding absolute frequencies are given below and above the bar.

Table 2.5: Proteins translated from hypothetical transcripts and genes predicted by the Ensembl annotation pipeline with unambiguous peptide evidence identified in the HUPO BPP pilot study.

IPI ID	Ensembl transcript ID	Ensembl gene ID	InterPro family (F) InterPro domain (D) Ensembl family (EF)
IPI00551011	ENST00000347377	ENSG00000055163	Cytoplasmic fragile X mental retardation protein interacting protein (F)
IPI00399121	ENST00000297820	ENSG00000165121	Kinesin, motor region (D)
IPI00246022	ENST00000326308	ENSG00000176661	H+-transporting two-sector ATPase (D)
IPI00218084	ENST00000316258	ENSG00000178510	Eukaryotic initiation factor 5A (F)
IPI00178323	ENST00000326994	ENSG00000178759	FAD dependent oxidoreductase (F)
IPI00402141	ENST00000324925	ENSG00000181312	Heat shock protein Hsp70 (F)
IPI00377005	ENST00000330554	ENSG00000183022	Tropomyosin (F)
IPI00457363	ENST00000328175	ENSG00000183247	Somatomedin B (D)
IPI00050211	ENST00000327852	ENSG00000184200	Keratin, type I (F)
IPI00176698	ENST00000332498	ENSG00000184844	Cytochrome c, class I (F)
IPI00253411	ENST00000327600	ENSG00000185439	Alpha tubulin (F)
IPI00243603	ENST00000335008	ENSG00000186728	Pyruvate kinase (F)
IPI00478539	ENST00000360026	ENSG00000196157	RNA-binding region RNP-1 (D) Heterogeneous Nuclear Ribonucleoprotein (F)
IPI00514712	ENST00000359559	ENSG00000196448	Double-stranded RNA binding (F)
IPI00477548	ENST00000358493	ENSG00000196829	-
IPI00453476	ENST00000355965	ENSG00000197041	Phosphoglycerate/ bisphosphoglycerate mutase (F)
IPI00374021	ENST00000264132	ENSG00000197106	Sodium:neurotransmitter symporter (F)
IPI00398418	ENST00000359736	ENSG00000197799	Bipartite nuclear localization signal (F)
IPI00021196	ENST00000360887	ENSG00000198214	-
IPI00514841	ENST00000361329	ENSG00000198751	ATP synthase D chain, mitochondrial (F)
IPI00477086	ENST00000359418	ENSG00000187647	ATP/GTP-binding site motif A (P-loop) (D) Periaxin (EF)
IPI00411506	ENST00000343340	ENSG00000128422	Keratin, type I (F)

Table 2.6: Proteins translated from hypothetical transcripts of known genes predicted by the Ensembl annotation pipeline with unambiguous peptide evidence identified in the HUPO BPP pilot study.

IPI ID	Ensembl transcript ID	Ensembl gene ID	gene symbol
IPI00479722	ENST00000258800	ENSG00000092010	*PSME1*
IPI00480211	ENST00000360140	ENSG00000100911	*PSME2*
IPI00172579	ENST00000354775	ENSG00000143149	*ALDH9A1*
IPI00022082	ENST00000296873	ENSG00000164402	*SEPT8*
IPI00478733	ENST00000358910	ENSG00000196176	*HIST1H4A*
IPI00479267	ENST00000358771	ENSG00000198053	*PTPNS1*
IPI00478410	ENST00000356708	ENSG00000165629	*ATP5C1*
IPI00472047	ENST00000354731	ENSG00000068903	*SIRT2*

2.3.3.5 Tissue expression

Tissue specificity of identified proteins was assessed based on transcript annotations obtained from the Ensembl database.

Table 2.7 summarises the eVOC tissue annotation of 1493 transcripts that encode proteins identified in the pilot study. Ontology terms and their frequency are shown according to their hierarchical order in the ontology. The ontology is pruned to the top level terms for non-nervous system tissues and the peripheral nervous system. Annotations for central nervous system specific terms are shown in detail.

The majority of identifications were annotated as ubiquitously expressed across a wide range of tissues. Almost all identifications encoded by an Ensembl transcript and associated with an eVOC term were annotated as expressed in the nervous system (98.6%) and brain (97.6%) on messenger RNA (mRNA) level. Interestingly, while 89.4 % of transcripts were annotated as expressed in the cerebral cortex, only one transcript was annotated specifically as expressed in the temporal lobe.

Based on eVOC annotation, 60 transcripts associated with 55 identified IPI entries have not been reported as expressed in brain before. However, twelve of them were found annotated as expressed in other parts of the nervous system. Only five proteins were annotated as nervous system specific, i.e. not expressed in any other tissue types (see table 2.8).

Table 2.7: Tissue expression of identified proteins based on eVOC annotation of the respective Ensembl transcripts. Shown is the frequency of transcripts annotated with each term. For non-nervous system tissues and the peripheral nervous system only the the frequency of the top level term is shown. Terms appear in the order of their position in the ontology hierarchy.

eVOC ID	eVOC term	freq.	freq. [%]
EV:0100000	Anatomical System	1493	100.0
EV:0100017	cardiovascular system	1313	87.9
EV:0100036	respiratory system	1442	96.6
EV:0100048	lymphoreticular system	1337	89.6
EV:0100056	alimentary system	1425	95.4
EV:0100088	liver and biliary system	1345	90.1
EV:0100094	urogenital system	1343	90.0
EV:0100128	endocrine system	1398	93.6
EV:0100139	musculoskeletal system	1318	88.3
EV:0100151	dermal system	1336	89.5

continued on next page

Table 2.7 – continued from previous page

eVOC ID	eVOC term	freq.	freq.
			[%]
EV:0100162	nervous system	1472	98.6
EV:0100335	peripheral nervous system	1378	92.3
EV:0100163	central nervous system	1462	97.9
EV:0100164	brain	1461	97.9
EV:0100165	cerebrum	1344	90.0
EV:0100166	cerebral cortex	1334	89.4
EV:0100180	hippocampus	626	41.9
EV:0100167	frontal lobe	951	63.7
EV:0100169	temporal lobe	1	0.1
EV:0100182	basal nuclei	1069	71.6
EV:0100184	corpus striatum	1	0.1
EV:0100185	caudate nucleus	1032	69.1
EV:0100189	amygdala	181	12.1
EV:0100194	diencephalon	1187	79.5
EV:0100195	thalamus	1020	68.3
EV:0100220	epithalamus	157	10.5
EV:0100221	pineal body	157	10.5
EV:0100223	subthalamus	8	0.5
EV:0100224	subthalamic nucleus	8	0.5
EV:0100225	hypothalamus	678	45.4
EV:0100241	brain stem	223	14.9
EV:0100242	midbrain	33	2.2
EV:0100247	substantia nigra	33	2.2
EV:0100253	pons	1	0.1
EV:0100271	trigeminal nucleus	1	0.1
EV:0100273	motor	1	0.1
EV:0100275	medulla oblongata	192	12.9

continued on next page

Table 2.7 – continued from previous page

eVOC ID	eVOC term	freq.	freq. [%]
EV:0100293	cerebellum	1089	72.9
EV:0100294	cerebellum cortex	136	9.1
EV:0100304	tract	1010	67.6
EV:0100305	corpus callosum	1010	67.6
EV:0100312	meninges	125	8.4
EV:0100313	dura mater	125	8.4
EV:0100315	pia matery	123	8.2
EV:0100316	spinal cord	1043	69.9

Table 2.8: Proteins identifications with nervous system specific eVOC annotation. Identifications associated with Ensembl transcripts annotated with the eVOC term *nervous system* or its descendent-terms but not with any other term of the eVOC *anatomical system* ontology.

IPI ID	description	gene symbol
IPI00010452	Transcriptional repressor scratch 1	*SCRT1*
IPI00154618	Zinc finger protein 312	*FEZF2*
IPI00240510	Transcription factor SOX-3	*SOX3*
IPI00409587	Myelin oligodendrocyte glycoprotein	*MOG*
IPI00514019	Myelin oligodendrocyte glycoprotein	*MOG*
IPI00410675	Syntaxin-1B2	*STX1B*

2.3.3.6 Correlation with transcript expression levels

A common problem in proteomics is the bias towards highly abundant and therefor often well known proteins as the results in section 2.3.3.1 indicate for the HUPO BPP dataset as well. To check if this bias relates to mRNA abundance levels and whether there is a correlation between the identifiability of proteins and their expression on transcript level, genome-wide distribution of expression levels was compared to that of genes identified in the HUPO BPP pilot study.

Su *et al.* have measured genome-wide mRNA expression level of protein-coding genes across a panel of 79 human tissues [229]. Figures 2.8 A and B show the distribution of expression levels measured in temporal lobe tissue for

1. all transcripts interrogated in the study by Su *et al.*
2. interrogated transcripts with a cross reference to Ensembl protein coding genes
3. the subset of Ensembl protein-coding genes identified in the HUPO BPP pilot study

The box plots in figure 2.8 A summarise the distribution of expression levels. The whiskers indicate the extreme (non-outlier) values observed, the borders of the boxes mark the lower and upper quartiles and the horizontal bar the location of the median. Figure 2.8 B shows the empirical cumulative distribution function (ECDF) for the three sets of transcripts.

The distribution of expression values was positively skewed, with the majority of transcripts having a low expression value. Although there was a trend towards higher expression levels observable for transcripts with cross references to Ensembl protein-coding genes in general, there was a much more pronounced shift towards higher expression levels for genes identified in the pilot study. The mean expression for Ensembl protein-coding genes was 1056 and the median 127, compared to 3021 and 322 respectively for the HUPO BPP set.

Figures 2.8 C and D show the same plots for genes identified in the pilot study, grouped by the identification frequency of their protein products.

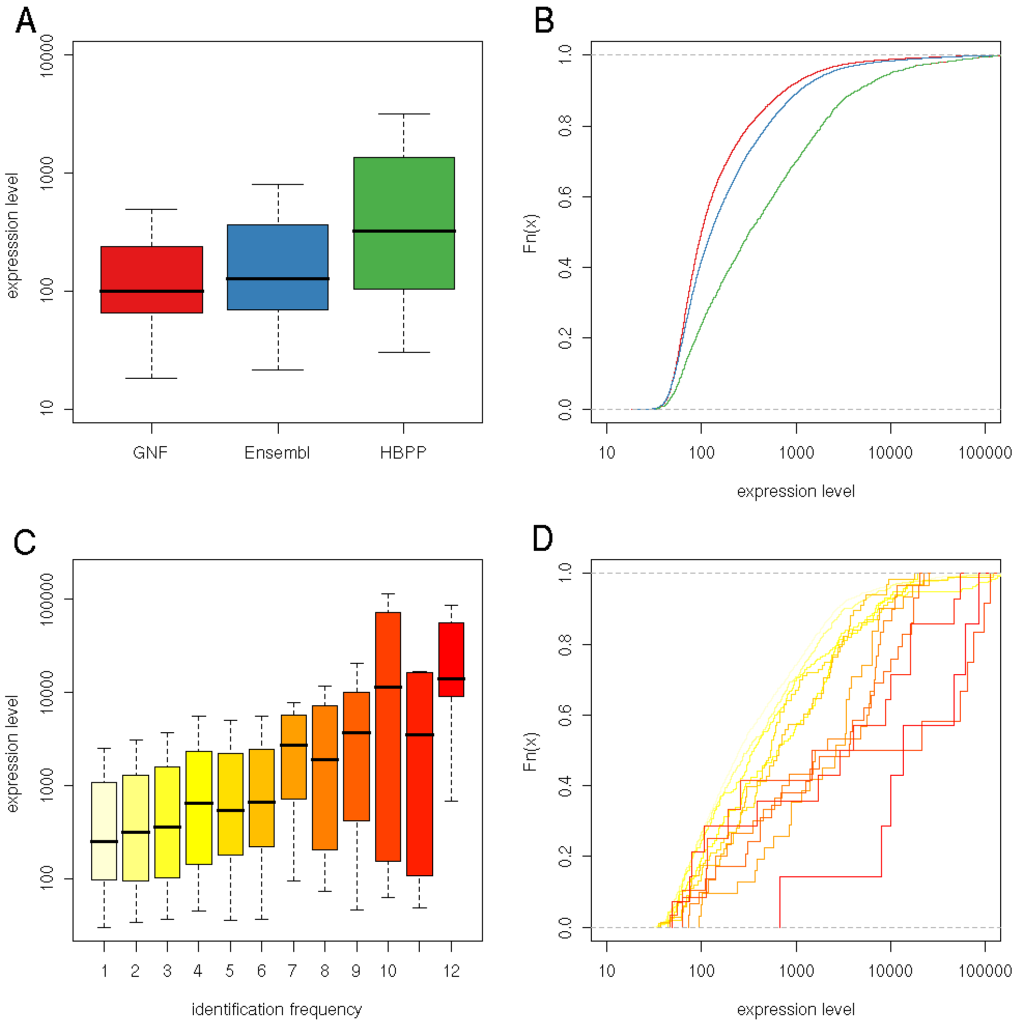


Figure 2.8: Transcript expression levels of proteins identified in the HUPO BPP pilot study. A) box plots summarising the distribution of mRNA expression levels measured in temporal lobe tissue for all transcripts in the GNF SymAtlas dataset (GNF) and the subsets of transcripts with cross references to Ensembl protein-coding genes (Ensembl) and Ensembl protein-coding genes identified in the HUPO BPP pilot study (HBPP). B) Plot of the empirical cumulative distribution function (ECDF) of mRNA expression levels for the respective transcript sets. C) box plots of mRNA expression levels of genes identified in the pilot study grouped by the identification frequency of their protein products. D) ECDF of the respective distributions. The colour coding of ECDF plots is the same as for the box plots.

The plots indicate a positive correlation between mRNA expression levels and the identification frequency of protein products (Spearman's rank correlation coefficient $r_s = 0.97$).

2.3.3.7 Subcellular localisation

To obtain an overview of the subcellular origin of proteins identified in the HUPO BPP pilot study, their sequences were analysed with regard to localisation signals and transmembrane domains. An overview of the functional composition of the dataset was obtained by categorising the proteins by functional annotation, protein family and biological pathway.

The putative subcellular origin of proteins was assessed based on the presence of localisation and transmembrane regions. The frequency of these sequence features in the HUPO BPP dataset is summarised in table 2.9.

Transit sequences triggering mitochondrial import were found in 125 sequences, of which 12 had transmembrane domains. An amino (N)-terminal signal sequence targeting the protein to the secretory pathway was featured in 124 proteins. Fourty seven proteins carrying a signal sequence were predicted transmembrane proteins, while 77 had no predicted transmembrane domains and were therefore likely to be secreted or to function in the endomembrane system. Another 141 proteins were predicted to be transmembrane proteins. The remaining 1416 identified proteins had neither localisation signals nor transmembrane domains, and were most likely of cytoplasmic origin.

Table 2.9: Localisation signals and transmembrane (TM) features of identified proteins.

		TM domain		
		yes	no	
Localisation Sequence	signal	47	77	124
	transit	12	113	125
	none	141	1416	1557
		200	1606	1806

In total, 200 proteins in the dataset were transmembrane proteins, corresponding to 11% of all identifications. A comparison with the relative frequency of transmembrane proteins in the Ensembl database (see figure 2.9 A) shows that the proportion of identified transmembrane domains was 50% lower than to be expected from the genome-wide frequency.

Also shown in figure 2.9 A is the frequency of transmembrane proteins for the different separation techniques employed in the pilot study. The frequency in LC and 1D-PAGE based experiments was similar to the overall frequency in the HUPO BPP study while only 3.4% of proteins identified in 2D-PAGE gels were transmembrane proteins.

Figure 2.9 B summarises the transmembrane domain frequency of proteins in the respective sets. A comparison of the distributions indicates a bias towards proteins with smaller numbers of transmembrane domains in the HUPO BPP data. Categorisation by separation method shows that gel-based methods are the major contributors to this trend.

A detailed view of the transmembrane domain frequency of identifications in comparison to Ensembl is shown in figure 2.9 C. Interestingly, proteins with seven transmembrane domains, which includes the family of G-protein coupled receptors (GPCR) important for signal transduction across the plasma membrane, are clearly under-represented in the study data.

A more detailed picture of the subcellular localisation was obtained from GO annotations. The GO Slim [341] distribution of HUPO BPP identifications vs all IPI entries for the GO *cellular component* ontology is shown in figure 2.10. GO Slim is a pruned down version of the ontology tree giving a broader overview of the GO classification.

Over- and under-representation of ontology terms was assessed based on the hypergeometric P value of the observed term frequency given its frequency in the IPI database. At the chosen FDR of 0.01% terms with $P_O \leq 1.8 \cdot 10^{-5}$ were considered as significantly over-represented

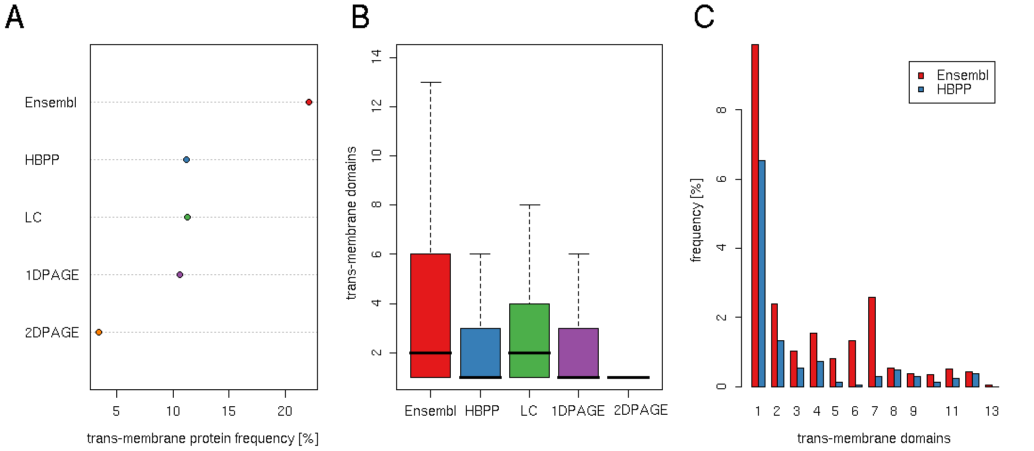


Figure 2.9: Identification of transmembrane proteins in the HUPO BPP pilot study. A) Dot plot comparing the relative frequency of transmembrane proteins in Ensembl, the HUPO BPP dataset and HUPO BPP identifications grouped by separation technique. B) Box plot summarising the transmembrane domain frequency of proteins in the respective sets. C) Bar chart showing the frequency of transmembrane proteins with one to thirteen transmembrane domains in Ensembl and the HUPO BPP dataset. All three plots are based on protein identifications with cross references to Ensembl.

and terms with $P_U \leq 7.6 \cdot 10^{-7}$ were considered as significantly under-represented. Frequency ratios of enriched terms are shown in green, frequency ratios of depleted terms in red in figure 2.10.

The analysis indicated a significant enrichment of proteins categorised as intracellular. Amongst these, cytoplasmic proteins were significantly over-represented ($P_O = 2.7 \cdot 10^{-112}$) with a frequency of 42.1% compared to 18.4% in IPI. Nuclear proteins, on the other hand, were significantly under-represented ($P_U = 5.7 \cdot 10^{-15}$) .

Given the low frequency of transmembrane proteins, one would expect the category *membrane* to be under-represented as well. That this was not the case is explained by the fact that this category includes peripheral membrane as well as integral membrane proteins, with only the latter being significantly under-represented ($P_U = 4.0 \cdot 10^{-23}$). As these terms are not included in GO Slim they are not shown in figure 2.10.

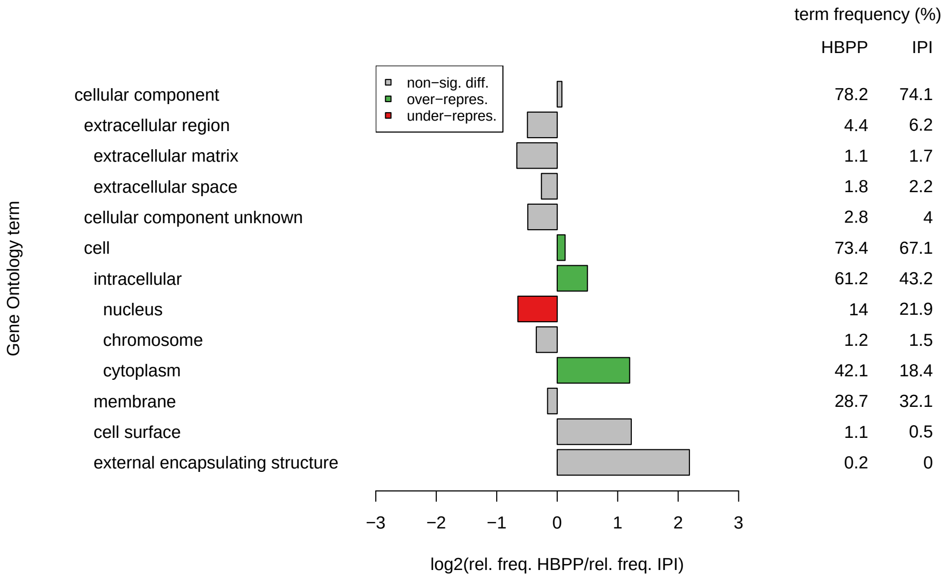


Figure 2.10: Categorisation of proteins identified in the HUPO BPP pilot by molecular function. Distribution of identified proteins across the GO Slim version of the *cellular component* ontology based on annotations obtained from the GOA database.The difference in term frequency is shown as the log_2 ratio of relative term frequency in the HUPO BPP dataset (HBPP) and the subset of IPI annotated in GOA (IPI). Significantly over-represented terms are highlighted in green, terms significantly under-represented are highlighted in red. sig. = significant, repres. = represented, diff. = difference.

Looking at the fine grained GO annotation of proteins categorised as cytoplasmic showed that the largest subset localises to intracellular organelles (717 identifications). Amongst these cytoplasmic membrane-bound vesicle components (51 proteins) and more specifically clathrin coated vesicles (35 proteins) - which includes synaptic vesicles (17 proteins) - were significantly enriched ($P_O = 9.9 \cdot 10^{-17}$). The second largest subgroup of identifications annotated with the GO term *cytoplasm* (215 proteins) were proteins associated with the cytoskeleton. Significantly enriched were components of the actin ($P_O = 6.4 \cdot 10^{-25}$) and microtubule cytoskeleton ($P_O = 3.0 \cdot 10^{-17}$), as well as neurofilaments ($P_O = 1.6 \cdot 10^{-5}$) which belong to the family of intermediate filaments.

The third group of significantly enriched cytoplasmic proteins was associated with mitochondria ($P_O = 2.6 \cdot 10^{-65}$). This includes peripheral inner membrane proteins that are part of the mitochondrial electron transport chain ($P_O = 2.4 \cdot 10^{-8}$) and proteins of the mitochondrial matrix ($P_O = 1.6 \cdot 10^{-7}$).

Enriched neuron-specific categories included *axon* ($P_O = 4.0 \cdot 10^{-8}$) and *growth cone* ($P_O = 3.4 \cdot 10^{-8}$) as well as *synaptosome* ($P_O = 6.2 \cdot 10^{-6}$).

2.3.3.8 Protein families

Protein families annotated in InterPro were ranked according to the number of family members identified in human brain specimens and compared to the ranking of protein families in IPI. The top ten superfamilies identified in the HUPO BPP study are shown in table 2.10.

The most abundant protein family in the HUPO BPP set was the superfamily of Ras GTPases with representatives of the Ran, Ras, Rho and Rab subfamilies present in the dataset. In IPI the Ras family was also found amongst the highest ranking families at position four.

The small GTPases were followed by families of cytoskeleton-constituting and -associated proteins including the tubulin, intermediate filament, actin and tropomyosin families.

Also found amongst the most frequent protein families were cation transporters mostly represented by members of the ATPase α-subunit subfamily and the calcium transporter subfamily.

While G-protein coupled receptor (GPCR)s were the most frequent protein family in IPI they were almost absent in the HUPO BPP set. This is consistent with the under-representation of transmembrane proteins in general and more specifically the group of seven transmembrane domain containing proteins mentioned above. The intracellular receptor components, the G-protein α-subunits, on the other hand, ranked relatively high at position nine. The remaining positions were occupied by the families of dynamin, chaperonin and heat shock proteins.

Table 2.10: Ranking of protein families identified in the HUPO BPP pilot study in comparison to the ranking in IPI. The list shows the top ten protein superfamilies identified in the HUPO BPP. The list is ordered by the frequency of the superfamilies in the HUPO BPP dataset (HBPP). Subfamilies are shown below in the order they appear in the InterPro hierarchy of protein families.

rank		InterPro	description	frequency	
HBPP	IPI	ID		HBPP	IPI
1	4	IPR001806	Ras GTPase	56	310
		IPR006689	ARF/SAR superfamily	5	59
		IPR006688	ADP-ribosylation factor	4	35
		IPR006687	GTP-binding protein SAR1	1	8
		IPR013753	Ras	47	183
		IPR002041	GTP-binding nuclear protein Ran	5	20
		IPR003577	Ras small GTPase, Ras type	16	83
		IPR003578	Ras small GTPase, Rho type	13	75
		IPR003579	Ras small GTPase, Rab type	30	120
2	19	IPR000217	Tubulin	24	100
		IPR002452	α-tubulin	11	27
		IPR002453	β-tubulin	13	57
3	9	IPR001664	Intermediate filament protein	22	169
		IPR002957	Keratin, type I	12	113
		IPR003054	Keratin, type II	2	48
4	11	IPR001757	ATPase, P-type, K/Mg/Cd/Cu/Zn/Na/ Ca/Na/H-transporter	17	149
		IPR000695	H+ transporting ATPase, proton pump	1	22
		IPR006069	ATPase, P-type cation exchange, α-subunit	8	24
		IPR005782	Calcium ATPase	1	12
		IPR006408	Calcium-translocating P-type ATPase, PMCA-type	10	28
		IPR006539	Phospholipid-translocating P-type ATPase, flippase	1	28
5	20	IPR004000	Actin/actin-like	15	92
		IPR004001	Actin	11	28
6	70	IPR001401	Dynamin	15	38
7	31	IPR000533	Tropomyosin	14	61
8	62	IPR002423	Chaperonin Cpn60/TCP-1	13	42
		IPR001844	Chaperonin Cpn60	8	18
		IPR002194	Chaperonin TCP-1	11	23
9	47	IPR001019	G-protein, α-subunit	12	54
10	82	IPR001023	Heat shock protein Hsp70	12	35
		IPR013126	Heat shock protein 70	11	21
		IPR012725	Chaperone DnaK	1	1

2.3.3.9 Functional categorisation

To obtain a more detailed picture of the functional composition of the HUPO BPP dataset, proteins were categorised by their functional annotation in the GOA database. Figures 2.11 and 2.12 show an overview of the functional categories of the HUPO BPP dataset based on their distribution across the GO Slim versions of the GO *molecular activity* and *biological process* ontologies.

Categorisation by *molecular function* showed a significant over-representation of proteins with catalytic ($P_O = 1.5 \cdot 10^{-24}$), structural molecule ($P_O = 8.3 \cdot 10^{-19}$), transporter ($P_O = 7.2 \cdot 10^{-15}$), binding ($P_O = 5.2 \cdot 10^{-18}$) and antioxidant ($P_O = 4.5 \cdot 10^{-6}$) activity.

Amongst over-represented catalytic activities are isomerase ($P_O = 8.7 \cdot 10^{-6}$), oxidoreductase ($P_O = 4.6 \cdot 10^{-24}$) and hydrolase ($P_O = 4.3 \cdot 10^{-12}$) activity. Significantly enriched enzyme activities falling into the latter two categories are *NADH dehydrogenase activity* ($P_O = 6.1 \cdot 10^{-17}$) and *ATPase activity coupled to transmembrane movement of ions* ($P_O = 6.1 \cdot 10^{-17}$) respectively. These terms are not shown in figure 2.11 as they are not part of the GO Slim.

The abundance of structural proteins is twice as high as to be expected from their frequency in IPI which is in accordance with the large proportion of cytoskeleton and cytoskeleton-association protein families seen earlier.

While ion and electron transporter activity ($P_O = 2.7 \cdot 10^{-8}$) as well as carrier activity ($P_O = 8.3 \cdot 10^{-24}$) are significantly enriched, transport activities associated with integral-membrane proteins, i.e. channel and pore class transporters ($P_O = 7.6 \cdot 10^{-7}$), are significantly under-represented.

The low frequency of transmembrane proteins and proteins of nuclear origin is reflected in the significant under-representation of functional categories associated with these cellular components. For instance the frequency of proteins with signal transducer activity ($P_U = 3.1 \cdot 10^{-26}$) is at 8.2% , half of that to be expected from the frequency in IPI (17.6%), and receptor activity is five-fold under-represented ($P_O = 3.5 \cdot 10^{-46}$) compared to IPI (2.2%). Significantly under-represented categories associated with nuclear function are *nucleic acid binding* ($P_U = 2.4 \cdot 10^{-18}$) and *transcription regulator activity* ($P_U = 2.4 \cdot 10^{-21}$).

The biases observed in the functional composition of the protein set translate into biases in the categorisation by *biological process* shown in figure 2.12. That is, proteins functioning in metabolism are significantly over-represented ($P_O = 2.5 \cdot 10^{-11}$), as are proteins involved in transport ($P_O = 3.9 \cdot 10^{-33}$), more specifically electron transport ($P_O = 4.2 \cdot 10^{-8}$). Also over-represented is the category *cell motility* ($P_O = 2.5 \cdot 10^{-10}$). Under-represented categories are *transcription* ($P_U = 1.1 \cdot 10^{-27}$), *regulation of transcription* ($P_U = 2.7 \cdot 10^{-27}$), and *regulation of metabolism* ($P_U = 1.3 \cdot 10^{-14}$).

2.3.3.10 Biological pathways

Using the Reactome database 216 identified proteins could be mapped to biological pathways [329]. The 1639 reactions and 704 pathways described in the database are grouped into 23 topics, of which 19 are labelled in a graphical overview of Reactome pathways as shown in figure 2.13.

Protein components of all pathway topics, apart from the notch signalling pathway and post-translational protein modification pathway, were identified. The highest coverage was obtained for proteins involved in electron transport (57.3%), and oxidative decarboxylation of pyruvate and the tricarboxylic acid (TCA) cycle (76.2%).

Proteins identified recurrently across several laboratories were involved in glucose, amino acid, lipid, and nucleotide metabolism, as well as oxidative decarboxylation of pyruvate and the TCA cycle. The presence of several proteins functioning in hemostasis could be a result of blood contamination.

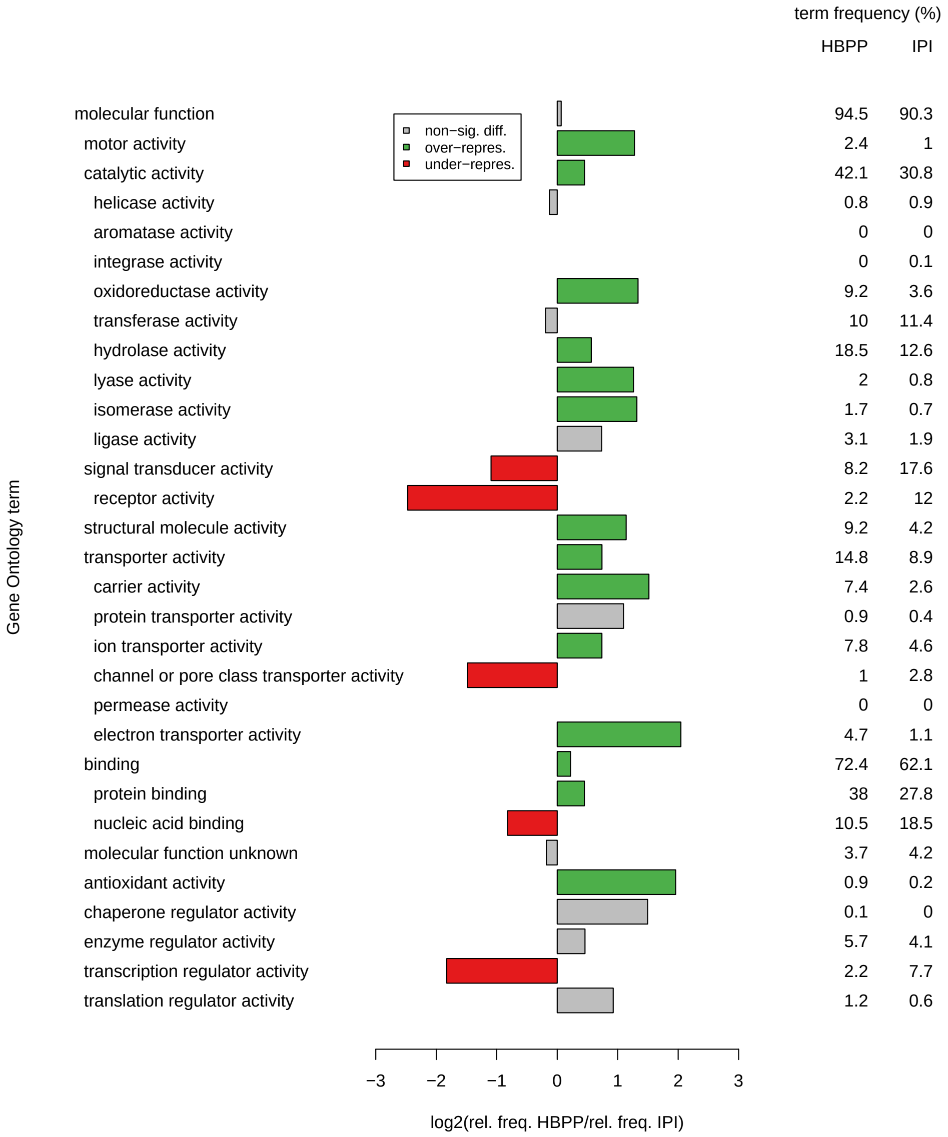


Figure 2.11: Categorisation of proteins identified in the HUPO BPP pilot by molecular function. Distribution of identified proteins across the GO Slim version of the *molecular function* ontology based on annotations obtained from the GOA database.The difference in term frequency is shown as the log_2 ratio of relative term frequency in the HUPO BPP dataset (HBPP) and the subset of IPI annotated in GOA (IPI). Significantly over-represented terms are highlighted in green, terms significantly under-represented are highlighted in red. sig. = significant, repres. = represented, diff. = difference.

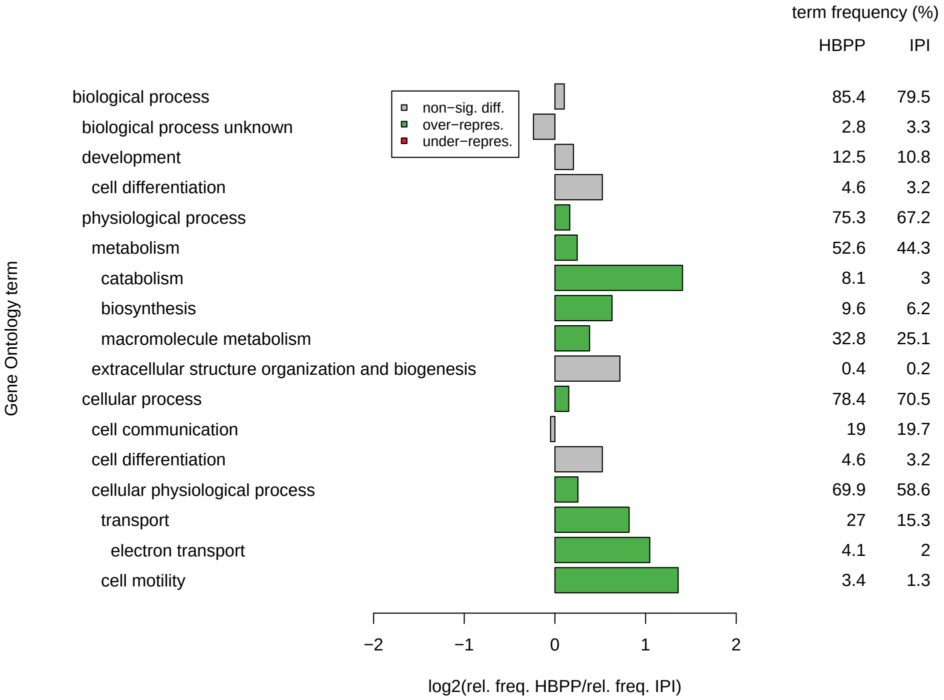


Figure 2.12: Categorisation of proteins identified in the HUPO BPP pilot study by biological process. Distribution of identified proteins across the GO Slim version of the *biological process* ontology based on annotations obtained from the GOA database. The difference in term frequency is shown as the log_2 ratio of relative term frequency in the HUPO BPP dataset (HBPP) and the subset of IPI annotated in GOA (IPI). Significantly over-represented terms are highlighted in green, terms significantly under-represented are highlighted in red. sig. = significant, repres. = represented, diff. = difference.

2.3.4 Detection of genes putatively pivotal to CNS function

So far the analysis had focused on the broader functional composition of the HUPO BPP dataset. In the context of the study of brain function, however, the most important aspect is the detection of gene products relevant to the study of brain or CNS specific function. To this end a list of genes likely to be pivotal to CNS function was compiled based on several

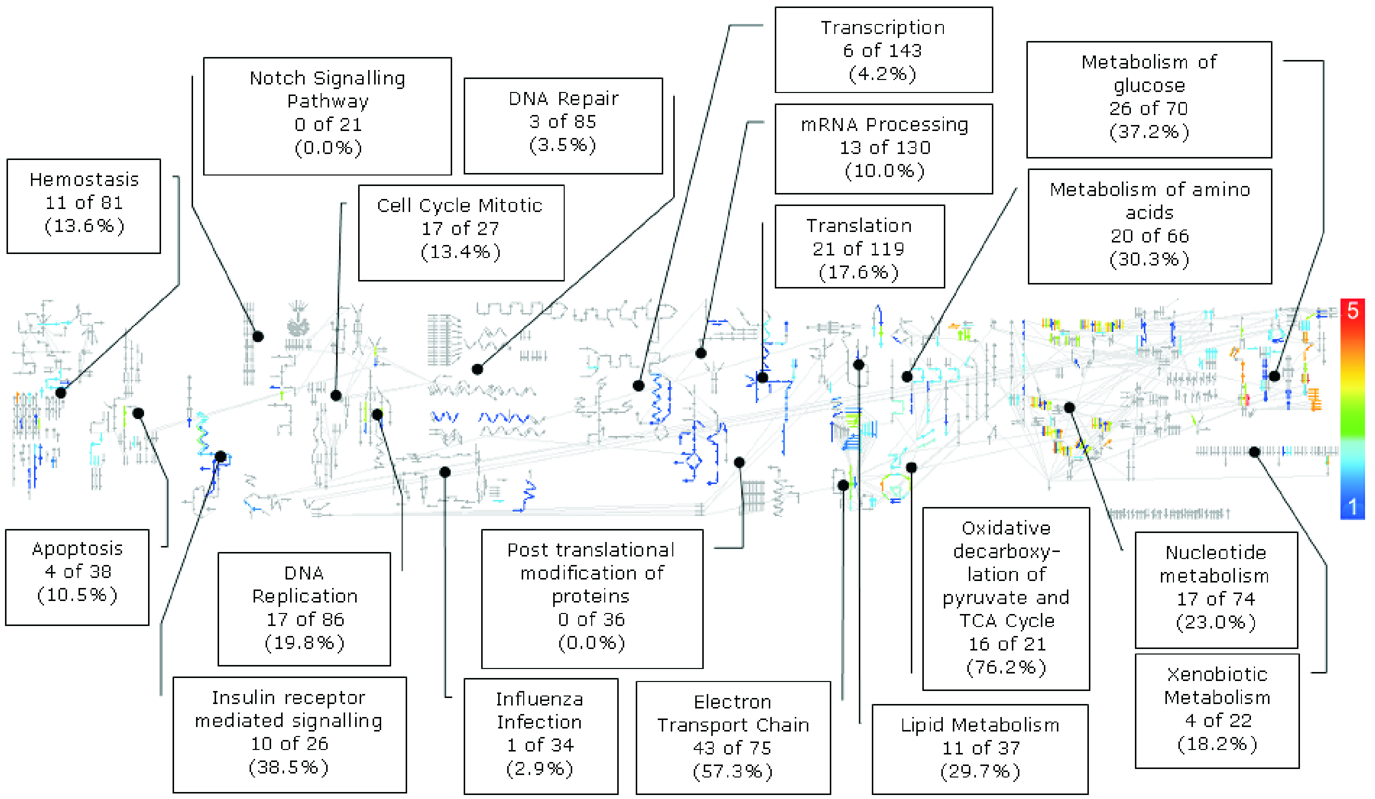


Figure 2.13: Distribution of identified proteins across biological pathways. Shown is an overview of the occurrence of protein identifications in pathways annotated in the Reactome database created with the Reactome SkyPainter tool. Pathways are depicted as a set of interconnected arrows, each representing a reaction. Arrows are coloured according to the number of laboratories that have identified proteins involved in the respective reaction: dark blue = 1, light blue = 2, green = 3, orange = 4, red = 5

lines of evidence. The coverage of genes in the list by peptide evidence in the HUPO BPP dataset was then assessed. Furthermore, peptide identifications from the HUPO BPP pilot study were compared to a large-scale study of protein expression in CSF.

2.3.4.1 Compilation of a list of genes putatively pivotal to CNS function

The list of CNS related genes was compiled based on the results of a PubMed search of the biomedical literature, curated annotations in the UniProtKB/Swiss-Prot database and mRNA expression profiles in the GNF SymAtlas of gene expression.

A non-redundant set of 418 (Ensembl) genes was identified that are either predominantly expressed in CNS or are thought to have a CNS specific splice isoform. These genes will be referred to as 'CNS related' genes for the remainder of the text.

The majority of genes (223) were selected on the basis of their expression profile, 149 genes occurred in CNS related publications, and the protein products of 83 genes were annotated as CNS specific or having a CNS specific splice isoform in the UniProtKB/Swiss-Prot database. The only gene identified by all three approaches encodes the microtubule-associated protein TAU. The origin of evidence for the 418 distinct genes is summarised in figure 2.14.

2.3.4.2 Construction of a maximal explanatory protein set for HUPO BPP peptide evidence

The minimal explanatory set of proteins reported originally emphasises the stringency of protein identification. A maximal explanatory set of proteins for the identified peptides, on the other hand, allows further elaboration on the specificity of peptide evidence on genome and isoform level. The 19308 distinct peptide identifications made across the nineteen experiments were therefore remapped to IPI version 3.29 to obtain a maximal

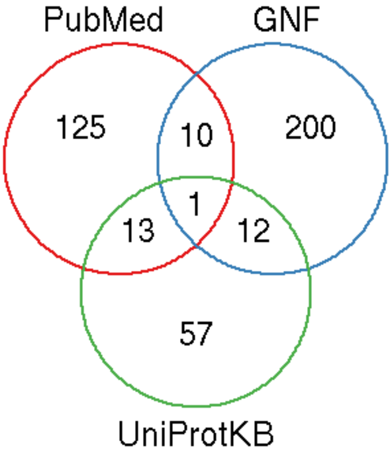


Figure 2.14: Evidence source for genes putatively pivotal to CNS function. Evidence for a gene being relevant to CNS function was obtained from a PubMed search of the biomedical literature (PubMed), curated annotation of proteins in the UniProtKB/Swiss-Prot (UniProtKB) database and mRNA expression profiles in the GNF SymAtlas of gene expression (GNF).

explanatory set of proteins for the peptide evidence. The remapping was based on exact string matching of HUPO BPP peptide sequences to IPI protein entries. The maximal explanatory set was constituted of all IPI entries matched by one or more HUPO BPP peptides.

Overall, 15451 peptides (80%) could be successfully mapped to entries in the more recent IPI version. The obtained set of proteins contained 9151 unique IPI accession numbers which correspondes to 13% of the 68161 entries in the IPI database. The set was around five times larger than the originally reported set. Mapping of the original set of IPI entries to IPI 3.29 using the IPI history file resulted in a set of 1737 IPI entries. Of these, 1718 proteins were contained in the set obtained by peptide remapping.

The nineteen mapped IPI entries without a peptide match corresponded to 61 proteins originally identified in IPI 3.05 by 415 distinct peptides. Of these peptides, 355 matched other proteins in the maximal explanatory set, including 197 peptide matches to 98 proteins mapped from the original set, and 158 matches to 958 proteins not reported previously.

To allow straightforward analysis of peptide evidence in the context of genome information from the Ensembl database, only those proteins with a cross reference to an Ensembl protein entry were retained from the initial set of mapped proteins. This reduced the number of proteins to 5941 (65%), a proportion slightly higher than the overall amount of IPI entries that have Ensembl cross references, which was the case for 35601 out of 68161 IPI proteins (52%). The set of corresponding Ensembl proteins contained 6193 Ensembl entries.

2.3.4.3 Peptide evidence for protein products of genes identified as putatively pivotal to CNS function

After further pruning the set of identified proteins with at least one Ensembl cross reference, to isolate only the protein products of the 418 CNS related Ensembl genes, a list of 395 Ensembl proteins corresponding to 388 IPI entries remained. Peptide evidence for these proteins amounted to 1699 distinct peptide sequences.

These proteins represented translations from 174 of the CNS related genes, thus providing tentative expression evidence for 41% of the CNS genes of interest. To assess the significance of the observed overlap it was compared to the average overlap of a hundred thousand random samples of IPI entries (figure 2.15). The comparison showed that the observed overlap is with a difference of four standard deviations significantly higher than the average overlap of 27% (143 genes) to be expected by chance ($p_t = 3 \cdot 10^{-5}$).

The 174 CNS related genes with peptide evidence contained a subset of 118 genes that encode proteins from the original minimal explanatory set. The respective set of 145 IPI entries was cross referenced to 170 Ensembl translations of CNS related genes. There were no genes originally reported that were not covered by the maximal explanatory set.

The maximal explanatory set contained an additional 181 isoforms for 68 of the 118 genes that were also contained in the minimal explanatory

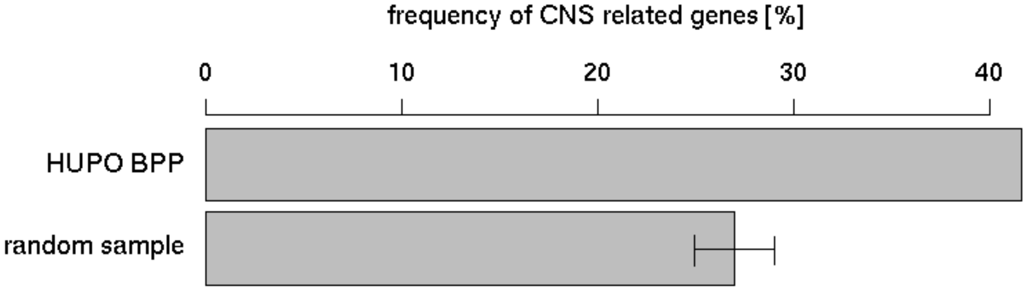


Figure 2.15: Peptide evidence for CNS related genes identified in the HUPO BPP. The bar plot shows the proportion of CNS related genes that encode proteins matched by peptides identified in the HUPO BPP pilot study (HUPO BPP) in comparison to the average overlap of 9151 IPI entries randomly sampled from IPI (random sample). The mean and standard deviation (indicated by the error bar) are based on 100000 trails.

set. Essentially these represented additional alternative hypotheses for the observed peptide evidence. A further 96 IPI entries represented protein products of the 56 CNS related genes not originally reported.

2.3.4.4 Specificity of peptide evidence on gene and splice isoform level

The majority of peptides matching CNS related genes (1402 [82%]) were associated with only one Ensembl gene. They identified 120 CNS related genes. The remaining 54 genes were matched by 297 peptides that matched multiple Ensembl genes from two to up to 351; 45 peptides with multiple matches mapped only to genes in the CNS gene set, while the remaining peptides also matched non-CNS related Ensembl genes.

If the stringency on peptide specificity was increased by requiring that gene-specific peptides must not match IPI entries not associated with a CNS related Ensembl gene, 548 peptides remained of which 542 were associated with exactly one gene. These peptides identified 86 genes with high confidence. Another six peptides were associated with two to three Ensembl genes, which identified an additional four genes.

Of the 86 genes with unambiguous peptide matches, 56 encoded alternative protein products ranging from two to seven splice isoforms. Figure 2.16 summarises the specificity of peptide evidence on splice isoform level.

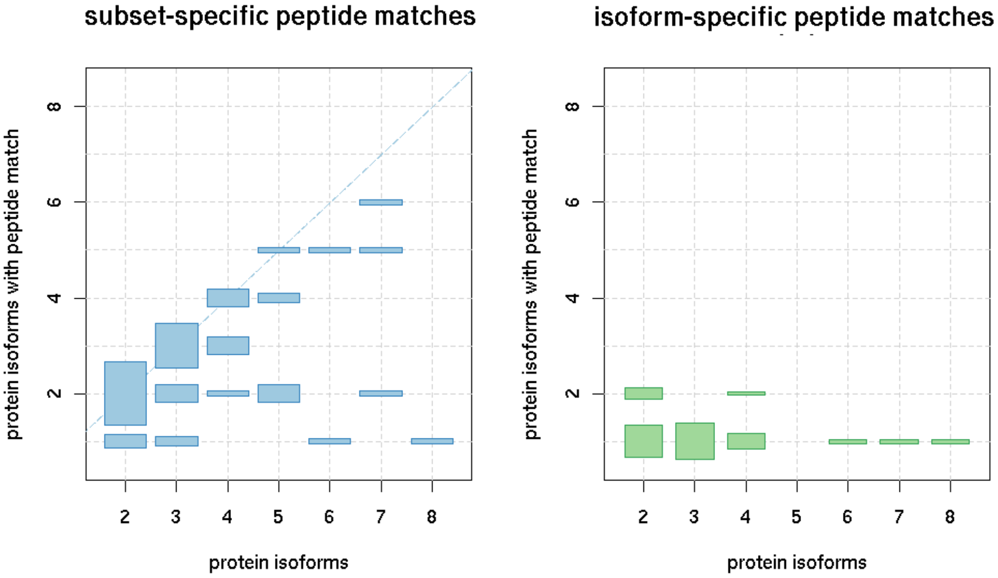


Figure 2.16: Splice isoform resolution of CNS related genes. Shown is the number of alternative protein products encoded by a gene on the x-axis versus the number of protein products with peptide matches on the y-axis. The height of the bar is proportional to the number of cases observed for each case. The plot on the left shows the proportion of protein products matched by non-isoform-specific and isoform-subset-specific peptides. The cases along the diagonal line have indiscriminate peptide matches across all isoforms, cases below the line have peptide matches specific to a subset of isoforms. The plot on the right shows the proportion of splice isoforms unambiguously identified by isoform specific peptide matches.

The plot on the left shows isoform resolution by ambiguous peptides, i.e. peptides that were not unique to a specific splice isoform, that could in some cases, however, discriminate subsets of isoforms. This was the case for 27 genes (table 2.11) that featured peptide evidence matching a subset of splice isoforms while evidence for the remaining splice isoforms was absent (cases below the dotted diagonal line). For another 29 alternatively spliced genes (table 2.12) identified by ambiguous peptide matches the evidence matched indiscriminately across all splice isoforms (cases along the dotted diagonal line).

Table 2.11: Alternatively spliced CNS genes with isoform-subset-specific peptide evidence. Given are the number of splice isoforms encoded by a gene and the number of isoforms matched by peptide evidence. IF = isoforms encoded by gene, IFID = isoforms with peptide match.

Ensembl Gene ID	gene symbol	description	IF	IF ID	IPI ID
ENSG00000103723	*AP3B2*	AP-3 complex subunit β-2	2	1	IPI00005793
ENSG00000145920	*CPLX2*	Complexin-2	2	1	IPI00012759
ENSG00000008056	*SYN1*	Synapsin-1	2	1	IPI00300568
ENSG00000046653	*GPM6B*	Neuronal membrane glycoprotein M6-b	3	1	IPI00187158
ENSG00000107902	*LHPP*	Phospholysine phosphohistidine inorganic pyrophosphate phosphatase	3	1	IPI00005474
ENSG00000033122	*LRRC7*	Leucine-rich repeat-containing protein 7	3	2	IPI00426267, IPI00426269
ENSG00000100167	*SEPT3*	Neuronal-specific septin-3	3	2	IPI00745056, IPI00384187
ENSG00000132639	*SNAP25*	Synaptosomal-associated protein 25	3	2	IPI00107625, IPI00010470
ENSG00000159082	*SYNJ1*	Synaptojanin-1	3	2	IPI00012441, IPI00333134
ENSG00000162706	*CADM3*	Cell adhesion molecule 3 precursor	4	2	IPI00009619, IPI00166048
ENSG00000163539	*CLASP2*	CLIP-associating protein 2	4	3	IPI00168165, IPI00024382, IPI00478227
ENSG00000133083	*DCAMKL1*	Serine/threonine-protein kinase DCLK1	4	3	IPI00004560, IPI00217247, IPI00644293
ENSG00000182621	*PLCB1*	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase β-1	4	3	IPI00219563, IPI00216920, IPI00395561

continued on next page

Table 2.11 – continued from previous page

Ensembl gene ID	gene symbol	description	IF	IF ID	IPI ID
ENSG00000065609	*SNAP91*	Synaptosomal-associated protein 91	4	3	IPI00646376, IPI00006612, IPI00470905
ENSG00000154027	*AK5*	Adenylate kinase isoenzyme 5	5	3	IPI00376041, IPI00844054, IPI00743623
ENSG00000206456	*AL669813.6*	Myelin-oligodendrocyte glycoprotein precursor	5	2	IPI00556079, IPI00744214
ENSG00000151150	*ANK3*	Ankyrin-3	5	2	IPI00333558, IPI00472779
ENSG00000116254	*CHD5*	Chromodomain-helicase-DNA-binding protein 5	5	3	IPI00646871, IPI00152535, IPI00445286
ENSG00000197971	*MBP*	Myelin basic protein	5	2	IPI00607642, IPI00216475
ENSG00000087470	*DNM1L*	Dynamin-1-like protein	6	5	IPI00146935, IPI00555883, IPI00235412, IPI00473085, IPI00037283
ENSG00000164402	*SEPT8*	Septin-8	6	1	IPI00549434
ENSG00000058404	*CAMK2B*	Calcium/calmodulin-dependent protein kinase type II β chain	7	4	IPI00334271, IPI00221305, IPI00183066, IPI00298090
ENSG00000126214	*KLC1*	Kinesin light chain 1	7	4	IPI00394906, IPI00784089, IPI00020096, IPI00337460
ENSG00000136153	*LMO7*	LIM domain only protein 7	7	5	IPI00291802, IPI00409591, IPI00552510, IPI00409593, IPI00409590

continued on next page

Table 2.11 – continued from previous page

Ensembl gene ID	gene symbol	description	IF	IF ID	IPI ID
ENSG00000186868	*MAPT*	Microtubule-associated protein tau	7	6	IPI00217976, IPI00025499, IPI00220175, IPI00747283, IPI00220174, IPI00026836
ENSG00000133318	*RTN3*	Reticulon-3	7	2	IPI00743293, IPI00398795
ENSG00000088367	*EPB41L1*	Band 4.1-like protein 1	8	1	IPI00384975

The plot on the right in figure 2.16 shows the proportion of splice isoforms identified by peptides which were unique across IPI and thus allowed unambiguous detection of a protein or splice isoform. CNS related proteins were identified by 229 such peptides matching 62 proteins encoded by 58 genes. However, 30 of these were non-alternatively spliced genes not included in the plot but listed in table 2.13.

The remaining 32 proteins were products of 28 genes that encoded between two and eight isoforms. They were identified by 96 unambiguous peptide matches. In all cases one or two isoforms have been resolved by specific peptide matches. For the three genes *GFAP*, *GNAO1* and *RTN1*, both known isoforms are resolved by protein-specific peptide sets. Furthermore, two of the four isoforms of *PSD3* are identified by one protein-specific peptide each. For the remaining 27 genes, one isoform is unambiguously identified by one or more protein-specific peptide matches.

2.3.4.5 Evidence for brain specific splice isoforms and isoforms previously without evidence on protein level

Details of splice isoforms identified by unique peptides are given in table 2.14. In some cases isoforms have been identified that are annotated as brain specific in the Swiss-Prot database.

Table 2.12: Alternatively spliced CNS genes with non-isoform-specific peptide matches. IF = isoforms encoded by gene.

Ensembl gene ID	gene Symbol	description	IF
ENSG00000160469	*BRSK1*	BR serine/threonine-protein kinase 1	2
ENSG00000135439	*CENTG1*	Centaurin-γ-1	2
ENSG00000175416	*CLTB*	Clathrin light chain B	2
ENSG00000169862	*CTNND2*	Catenin δ-2	2
ENSG00000187672	*ERC2*	ERC protein 2	2
ENSG00000131095	*GFAP*	Glial fibrillary acidic protein	2
ENSG00000087258	*GNAO1*	Guanine nucleotide-binding protein G(o) subunit α	2
ENSG00000176956	*LY6H*	Lymphocyte antigen 6H precursor	2
ENSG00000123560	*PLP1*	Myelin proteolipid protein	2
ENSG00000177380	*PPFIA3*	Liprin-alpha-3	2
ENSG00000131771	*PPP1R1B*	Protein phosphatase 1 regulatory subunit 1B	2
ENSG00000139970	*RTN1*	Reticulon-1	2
ENSG00000079215	*SLC1A3*	Excitatory amino acid transporter 1	2
ENSG00000130540	*SULT4A1*	Sulfotransferase 4A1	2
ENSG00000132692	*BCAN*	Brevican core protein precursor	3
ENSG00000106976	*DNM1*	Dynamin-1	3
ENSG00000197959	*DNM3*	Dynamin-3	3
ENSG00000067798	*NAV3*	Neuron navigator 3	3
ENSG00000104725	*NEFL*	Neurofilament light polypeptide	3
ENSG00000100804	*PSMB5*	Proteasome subunit β type-5 precursor	3
ENSG00000089169	*RPH3A*	Rabphilin-3A	3
ENSG00000198513	*SPG3A*	Atlastin-1	3
ENSG00000185666	*SYN3*	Synapsin-3	3
ENSG00000116147	*TNR*	Tenascin-R precursor	3
ENSG00000157087	*ATP2B2*	Plasma membrane calcium-transporting ATPase 2	4
ENSG00000164076	*CAMKV*	CaM kinase-like vesicle-associated protein	4
ENSG00000054523	*KIF1B*	Kinesin-like protein KIF1B	4
ENSG00000156011	*PSD3*	PH and SEC7 domain-containing protein 3	4
ENSG00000115310	*RTN4*	Reticulon-4	5

For example the unambiguously identified isoform PIKE-L of *CENTG1* and the *CLTB* isoform are brain specific. Unambiguously identified isoform CRA_e of *GPM6B* had only a TrEMBL entry and was not yet associated with the *GPM6B* Swiss-Prot entry. Furthermore, the TrEMBL entry states that there has only been evidence on transcript level for this protein. Another unambiguously identified splice isoform without previous experimental confirmation according to Swiss-Prot was isoform 3 of *CAMKV*.

Table 2.13: Non-alternatively spliced CNS genes with unambiguous peptide evidence.

Ensembl gene ID	gene symbol	description	IPI ID
ENSG00000130203	*APOE*	Apolipoprotein E precursor	IPI00021842
ENSG00000184524	*CEND1*	Cell cycle exit and neuronal differentiation protein 1	IPI00295601
ENSG00000047457	*CP*	Ceruloplasmin precursor	IPI00017601
ENSG00000100884	*CPNE6*	Copine-6	IPI00295469
ENSG00000132912	*DCTN4*	Dynactin subunit4	IPI00550852
ENSG00000175497	*DPP10*	Inactive dipeptidyl peptidase 10	IPI00464986
ENSG00000074800	*ENO1*	α-enolase	IPI00465248
ENSG00000111674	*ENO2*	γ-enolase	IPI00216171
ENSG00000108515	*ENO3*	β-enolase	IPI00218474
ENSG00000162188	*GNG3*	Guaninenucleotide-binding protein G(I)/G(S)/G(O) subunitγ-3 precursor	IPI00023625
ENSG00000132702	*HAPLN2*	Hyaluronan and proteoglycan link protein 2 precursor	IPI00029184
ENSG00000121905	*HPCA*	Neuron-specific calcium-binding protein hippocalcin	IPI00219103
ENSG00000116983	*HPCAL4*	Hippocalcin-like protein 4	IPI00008305
ENSG00000170049	*KCNAB3*	Voltage-gated potassium channel subunit β-3	IPI00006204
ENSG00000087250	*MT3*	Metallothionein-3	IPI00016666
ENSG00000117691	*NENF*	Neudesin precursor	IPI00002525
ENSG00000126861	*OMG*	Oligodendrocyte-myelin glycoprotein precursor	IPI00295832
ENSG00000168490	*PHYHIP*	Phytanoyl-CoA hydroxylase-interacting protein	IPI00022021
ENSG00000106278	*PTPRZ1*	Receptor-type tyrosine-protein phosphataseζ precursor	IPI00748312
ENSG00000100362	*PVALB*	Parvalbuminα	IPI00219703
ENSG00000160307	*S100B*	Protein S100-B	IPI00299399
ENSG00000170616	*SCRT1*	Transcriptional repressor scratch 1	IPI00010452
ENSG00000104969	*SGTA*	Smallglutamine-rich tetratricopeptide repeat-containing proteinα	IPI00013949
ENSG00000107295	*SH3GL2*	Endophilin-A1	IPI00019171
ENSG00000104888	*SLC17A7*	Vesicular glutamate transporter 1	IPI00025331
ENSG00000110436	*SLC1A2*	Excitatory amino acid transporter 2	IPI00300020
ENSG00000074317	*SNCB*	β-synuclein	IPI00032904
ENSG00000185518	*SV2B*	Synaptic vesicle glycoprotein 2B	IPI00006631
ENSG00000109654	*TRIM2*	Tripartite motif-containing protein 2	IPI00153011
ENSG00000163032	*VSNL1*	Visinin-like protein 1	IPI00216313

Table 2.14: Alternatively spliced CNS genes with isoform specific peptide evidence. Given are the number of splice isoforms encoded by a gene and the number of isoforms matched by unambiguou peptide evidence. IF = isoforms encoded by gene, IFID = isoforms with peptide match.

Ensembl gene ID	gene symbol	description	IF	IF ID	IPI ID	UniProtKB isoform ID	isoform name
ENSG00000103723	*AP3B2*	AP-3 complex subunit β-2	2	1	IPI00005793	Q13367	-
ENSG00000135439	*CENTG1*	Centaurin-γ-1	2	1	IPI00217393	Q99490-1	PIKE-L
ENSG00000175416	*CLTB*	Clathrin light chain B	2	1	IPI00014589	P09497-1	Isoform Brain
ENSG00000145920	*CPLX2*	Complexin-2	2	1	IPI00012759	Q6PUV4	-
ENSG00000187672	*ERC2*	ERC protein 2	2	1	IPI00472614	O15083-1	Isoform 1
ENSG00000131095	*GFAP*	Glial fibrillary acidic protein	2	2	IPI00025363, IPI00443478	P14136-1, P14136-3	GFAP α, GFAP ϵ
ENSG00000087258	*GNAO1*	Guanine nucleotide-binding protein G(o) subunit α	2	2	IPI00220281, IPI00398700	P09471-1, P09471-2	Alpha 1, Alpha 2
ENSG00000176956	*LY6H*	Lymphocyte antigen 6H precursor	2	1	IPI00014964	O94772	-
ENSG00000139970	*RTN1*	Reticulon-1	2	2	IPI00003971, IPI00219813	Q16799-1, Q16799-3	RTN1-A, RTN1-C
ENSG00000130540	*SULT4A1*	Sulfotransferase 4A1	2	1	IPI00161549	Q9BR01-1	Isoform 1
ENSG00000008056	*SYN1*	Synapsin-1	2	1	IPI00300568	P17600-1	Isoform IA
ENSG00000132692	*BCAN*	Brevican core protein precursor	3	1	IPI00456623	Q96GW7-1	Isoform 1
ENSG00000197959	*DNM3*	Dynamin-3	3	1	IPI00221332	Q9UQ16-1	Isoform 1

continued on next page

Table 2.14 – continued from previous page

Ensembl gene ID	gene symbol	description	IF	IF ID	IPI ID	UniProtKB isoform ID	isoform name
ENSG00000046653	*GPM6B*	Neuronal membrane glycoprotein M6-b	3	1	IPI00187158	Q8N956	Isoform CRA_e
ENSG00000107902	*LHPP*	Phospholysine phosphohistidine inorganic pyrophosphate phosphatase	3	1	IPI00005474	Q9H008	-
ENSG00000104725	*NEFL*	Neurofilament light polypeptide	3	1	IPI00237671	P07196	-
ENSG00000100167	*SEPT3*	Neuronal-specific septin-3	3	1	IPI00745056	Q9UH03-2	SEP3B
ENSG00000132639	*SNAP25*	Synaptosomal-associated protein 25	3	1	IPI00107625	P60880-2	Isoform SNAP-25a
ENSG00000185666	*SYN3*	Synapsin-3	3	1	IPI00298984	O14994	-
ENSG00000116147	*TNR*	Tenascin-R precursor	3	1	IPI00160552	Q92752-1	Isoform 1
ENSG00000164076	*CAMKV*	CaM kinase-like vesicle-associated protein	4	1	IPI00304600	Q8NCB2-3	Isoform 3
ENSG00000163539	*CLASP2*	CLIP-associating protein 2	4	1	IPI00024382	-	-
ENSG00000182621	*PLCB1*	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase β-1	4	1	IPI00219563	Q9NQ66-1	Isoform A
ENSG00000156011	*PSD3*	PH and SEC7 domain-containing protein 3	4	2	IPI00410060, IPI00166002	Q9NYI0-2, Q9NYI0-3	Isoform 2, Isoform 3
ENSG00000065609	*SNAP91*	Synaptosomal-associated protein 91	4	1	IPI00006612	O60641-1	Isoform 1
ENSG00000164402	*SEPT8*	Septin-8	6	1	IPI00549434	Q92599-1	SEPT8_v2
ENSG00000058404	*CAMK2B*	Calcium/calmodulin-dependent protein kinase type II beta chain	7	1	IPI00183066	Q13554-7	Beta7
ENSG00000088367	*EPB41L1*	Band 4.1-like protein 1	8	1	IPI00384975	-	-

2.3.4.6 Proteotypic character of protein-specific peptides identifying CNS related gene products

The protein-specific peptides identified are candidate proteotypic peptides which are defined as peptides unambiguously identifying a particular protein sequence and that are consistently and repeatedly identified by MS for any particular protein in a mixture. To assess the performance of proteotypic peptide predictions by Mallick and colleagues the experimintally identified protein-specific peptides were compared to the computationally predicted.

Of the 229 protein-specific peptides, 92 (40%) were predicted to be proteotypic for the protein they identify. Predictions included the separation and detection method a sequence is proteotypic for. Figure 2.17 A shows the overlap between the predictions and the observed peptide identifications against a background of random matching obtained by permutation testing. Fourty seven of the 92 peptides (51%) were identified at least once using one of the predicted separation methods, 62 peptides (67%) were detected by the predicted MS technology. Only 27 peptides, 29% of all protein-specific peptides, were observed by one of the predicted combinations of separation and detection method. Overall, the predictions did not perform better than a random assignment for this peptide population, as shown by the permutation test data.

As proteotypic peptides are required to be observed frequently, the protein-specific peptides were filtered by observation frequency only. By retaining those peptides whose identification frequency was above the mean frequency of all gene-specific peptides matching the protein sequence, a subset of 41 empirical proteotypic peptides was extracted. Table 2.15 shows the comparison between these 41 peptides and the corresponding proteotypic peptide predictions.

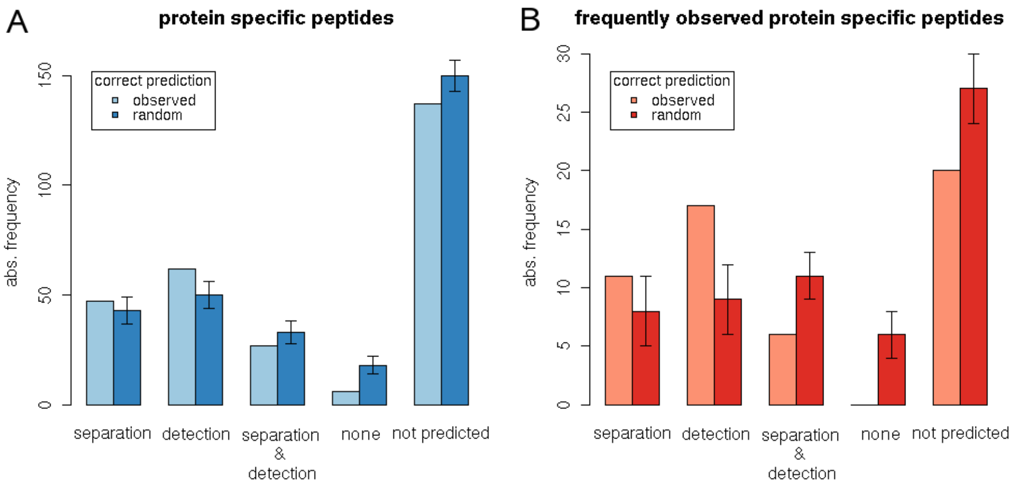


Figure 2.17: Performance of proteotypic peptide predictions for protein-specific peptides identifying CNS related. The bar charts show the overlap between the proteotypic peptide predictions and the observed data for separation and detection method for all the protein-specific peptides (A), and the frequently observed subset of these (B). The first two categories represent those cases in which the experimentally observed peptide was derived at least once from the predicted separation or detection method respectively. The third category requires that the peptide sequence was observed using the exact combination of the predicted separation and detection method, i.e. at least one experiment found the peptide using both the predicted separation method and the predicted detection method. 'None' means that the observed peptide was predicted to be proteotypic but neither separation nor detection matched the observation. The last category are observed protein-specific peptides that were not predicted to be proteotypic.

Table 2.15: Proteotypic peptide predictions for protein-specific peptides identifying CNS related gene products. The table lists for each experimentally observed peptide sequence the observation frequency, the average observation frequency of all peptides matching the protein identified by the peptide, the IPI ID of the identified protein, the respective Ensembl gene ID and gene symbol, the technologies the peptide was observed by and the proteotypic peptide sequence and technologies as predicted by Mallick *et al.* The score assigned to each prediction (on a scale between 0 and 1) is shown in brackets after the technology.

peptide sequence	observation frequency	average frequency	IPI ID	Ensembl gene ID	gene symbol	technology observation	proteotypic peptide prediction	technology prediction
LAPEVMEDLVK	3	2.8	IPI00216313	ENSG00000163032	*VSNL1*	PAGE_ESI LC_ESI	K.LAPEVMEDLVK.S	MUDPIT_ESI(>0.99)
GNPTVEVDLFTSK	8	3.8	IPI00465248	ENSG00000074800	*ENO1*	PAGE_ESI LC_ESI PAGE_MALDI PAGE/LC_ESI	R.GNPTVEVDLFTSK.G	MUDPIT_ESI(>0.99)
GNPTVEVDLHTAK	2	1.4	IPI00218474	ENSG00000108515	*ENO3*	PAGE_ESI PAGE_ESI	-	-

SQAIDLLYWR	2	1.5	IPI00219813	ENSG00000139970	*RTN1*	LC_ESI	K.SQAIDLLYWR.D	PAGE_ESI(0.91)
								PAGE_MALDI(>0.99)

STEFNEHELK	3	2.8	IPI00216313	ENSG00000163032	*VSNL1*	PAGE_ESI LC_ESI	-	-

SCCSCCPAECEK	4	2.5	IPI00016666	ENSG00000087250	*MT3*	LC_ESI	K.SCCSCCPAECEK.C	ICAT_ESI(>0.99)
								PAGE_ESI(0.98)
								PAGE_MALDI(0.96)

ELINNELSHFLEEIK	4	3	IPI00299399	ENSG00000160307	*S100B*	LC_ESI PAGE_ESI	E.LINNELSHFLEEIK.E	PAGE_MALDI(>0.99)

GATPAPQAGEPSPGLGAR	2	1.7	IPI00003971	ENSG00000139970	*RTN1*	PAGE_ESI
						LC_ESI
							K.GATPAPQAGEPSPGLGAR.A	PAGE_ESI(>0.99)
								MUDPIT_ESI(>0.99)

LNLEEFQQLYVK	5	2.8	IPI00216313	ENSG00000163032	*VSNL1*	PAGE/LC_ESI PAGE_ESI LC_ESI	R.LNLEEFQQLYVK.F	PAGE_MALDI(>0.99)

EGVVQGVASVAEK	6	3.6	IPI00032904	ENSG00000074317	*SNCB*	PAGE_ESI
						PAGE_MALDI
						LC_ESI
							R.EGVVQGVASVAEK.T	PAGE_ESI(>0.99)
								MUDPIT_ESI(>0.99)

GGEAAEAEAEK	4	2.5	IPI00016666	ENSG00000087250	*MT3*	LC_ESI	K.GGEAAEAEAEK.C	MUDPIT_ESI (>0.99)

continued on next page

Table 2.15 – continued from previous page

peptide sequence	observation frequency	average frequency	IPI ID	Ensembl gene ID	gene symbol	technology observation	proteotypic peptide prediction	technology prediction
TKEQASHLGGAVFSGAGNIAAATGLVK	4	3.6	IPI00032904	ENSG00000074317	*SNCB*	PAGE_ESI PAGE_MALDI LC_ESI	-	-
EQASHLGGAVFSGAGNIAAATGLVK	6	3.6	IPI00032904	ENSG00000074317	*SNCB*	LC_ESI PAGE_MALDI	-	-
AVDHINSTIAPALISSGLSVVEQEK	9	7.5	IPI00216171	ENSG00000111674	*ENO2*	LC_ESI PAGE_MALDI PAGE_ESI PAGE/LC_ESI	-	-
TIAPALVSK	7	3.8	IPI00465248	ENSG00000074800	*ENO1*	PAGE/LC_ESI PAGE_ESI PAGE_MALDI LC_ESI	-	-

AALEQPCEGSLTRPK	4	2.2	IPI00012759	ENSG00000145920	*CPL12*	LC_ESI
						PAGE_MALDI
							K.AALEQPCEGSLTR.P	PAGE_MALDI (0.99)
								PAGE_ESI(0.96)
								ICAT_ESI(>0.99)
								MUDPIT_ESI(>0.99)

SELEEQLTPVAEETR	2	1.3	IPI00021842	ENSG00000130203	*APOE*	PAGE_ESI	K.SELEEQLTPVAEETR.A	MUDPIT_ESI(>0.99)
AIPAGCGDEEEEEEESILDTVLK	3	2.2	IPI00012759	ENSG00000145920	*CPL12*	LC_ESI	K.AIPAGCGDEEEEEEESILDTVLK.Y	MUDPIT_ESI (>0.99)
LNWAFNMYDLDGDGK	4	2.8	IPI00216313	ENSG00000163032	VSNL1	PAGE_ESI LC_ESI	-	-
MNEDGLTPEQR	3	2.8	IPI00216313	ENSG00000163032	*VSNL1*	PAGE_ESI LC_ESI	-	-
GLFSSDSGIEMTPAESTEVNK	2	1.7	IPI00003971	ENSG00000139970	*RTN1*	PAGE_ESI LC_ESI	R.GLFSSDSGIEMTPAESTEVNK.I	MUDPIT_ESI(>0.99)
NKDDQITLDEFK	5	2.8	IPI00216313	ENSG00000163032	*VSNL1*	PAGE/LC_ESI PAGE_ESI LC_ESI	-	-
SDPSIVLLLQCDIQK	5	2.8	IPI00216313	ENSG00000163032	*VSNL1*	PAGE_ESI LC_ESI PAGE/ICAT_- MALDI/ESI	K.SDPSIVLLLQCDIQ.K	ICAT_ESI(>0.99)
YIDITRPEEVK	2	1.7	IPI00003971	ENSG00000139970	*RTN1*	PAGE_ESI	-	-

AATVGSLAGQPLQER	2	1.3	IPI00021842	ENSG00000130203	*APOE*	PAGE_ESI	R.AATVGSLAGQPLQER.A	PAGE_ESI(>0.99)
								MUDPIT_ESI(>0.99)

DTDISIKPEGVR	2	1.7	IPI00003971	ENSG00000139970	*RTN1*	PAGE_ESI LC_ESI	-	-

MHESLMLFDSICNNK	3	2.7	IPI00220281	ENSG00000087258	*GNAO1*	PAGE_ESI
						LC_ESI
							R.MHESLMLFDSICNNK.F	PAGE_MALDI(>0.99)
								ICAT_ESI(>0.99)

NGKYDLDFK	2	1.4	IPI00218474	ENSG00000108515	*ENO3*	PAGE_ESI LC_ESI	-	-

continued on next page

Table 2.15 – continued from previous page

peptide sequence	observation frequency	average frequency	IPI ID	Ensembl gene ID	gene symbol	technology observation	proteotypic peptide prediction	technology prediction
FFIDTSIILFLNK	3	2.7	IPI00220281	ENSG00000087258	*GNAO1*	PAGE_MALDI LC_ESI	-	-
AKLEEQAQQIR	2	1.3	IPI00021842	ENSG00000130203	*APOE*	PAGE_ESI	-	-
RAPQITTPVK	2	1.7	IPI00003971	ENSG00000139970	*RTN1*	PAGE_ESI	R.APQITTPVK.I	PAGE_ESI(>0.99)
ITLTEIEPSVETTTQEK	3	1.7	IPI00003971	ENSG00000139970	*RTN1*	PAGE_ESI LC_ESI	K.ITLTEIEPSVETTTQEK.T	MUDPIT_ESI(>0.99)
VNQIGSVTESIQACK	2	1.4	IPI00218474	ENSG00000108515	*ENO3*	PAGE_ESI LC_ESI	K.VNQIGSVTESIQACK.L	ICAT_ESI(>0.99)
GLSYETAENPRPVGQLADRPEVK	2	1.7	IPI00003971	ENSG00000139970	*RTN1*	PAGE_ESI	-	-
TVEMRDGEVIK	6	4	IPI00025363	ENSG00000131095	*GFAP*	PAGE/LC_ESI PAGE_ESI LC_ESI	-	-
KPAPPHPHLNK	2	1.4	IPI00298984	ENSG00000185666	*SYN3*	PAGE_ESI	-	-
MYGAHLASISTPEEQDFINNR	3	1.4	IPI00456623	ENSG00000132692	*BCAN*	PAGE_ESI LC_ESI	-	-
EYQWIGLNDR	2	1.4	IPI00456623	ENSG00000132692	*BCAN*	PAGE_ESI LC_ESI	R.EYQWIGLNDR.T	PAGE_MALDI(>0.99)
LILNYSPR	4	2	IPI00160552	ENSG00000116147	*TNR*	PAGE_ESI LC_ESI	R.LILNYSPR.D	PAGE_MALDI(>0.99)
WEAPQISCVPR	3	1.4	IPI00456623	ENSG00000132692	*BCAN*	PAGE/ICAT_- MALDI/ESI LC_ESI	-	-
LQVELEQEYQDK	2	1.4	IPI00219563	ENSG00000182621	*PLCB1*	PAGE_ESI	-	-

In slightly more than 50% of the cases (21 peptides), frequently observed peptides were predicted to be proteotypic and the peptides AATVGSLAGQPLQER (*APOE*), SELEEQLTPVAEETR (*APOE*), GNPTVEVDLFTSK (*ENO1*), and EGVVQGVASVAEK (*SNCB*) were observed at least once by the exact combination of predicted separation and detection method. However, a number of peptides reoccurring across many experiments were not predicted to be proteotypic at all. Figure 2.17 B indicates that the predictions for these 21 peptides performed slightly better than random, although the difference between random assignment and prediction was not significant ($\leq$ 1.5 times the standard deviation).

2.3.4.7 Comparison with of HUPO BPP and CSF peptide evidence for CNS related genes

The 4285 non-redundant peptide identifications made in CSF by Pan and colleagues were remapped to IPI version 3.29 in the same manner as for the HUPO BPP study resulting in a maximal explanatory set containing 2185 IPI entries of which 2103 had an Ensembl cross reference associating them with 828 Ensembl genes. The intersection of the HUPO BPP and CSF protein sets contained 1010 IPI entries associated with 378 genes. The maximal explanatory set contained expression evidence for 31 (7%) of the CNS related genes which was a considerably smaller proportion than covered by the HUPO BPP data.

The CSF set shared 24 CNS genes with the HUPO BPP set (table 2.16). The majority of proteins encoded by these genes carried an N-terminal signal sequence targeting them to the secretory pathway and three quarters of the genes with a signal sequence had transmembrane domains (table 2.16).

The CSF peptide evidence covered the same 58 protein products as the HUPO BPP evidence. However, only three peptides specifically identifying CNS related genes were shared between the two datasets: the peptide LAQANGWGVMVSHR identifying the protein product of *ENO1*, the peptide VATHLSTPQGLQFK matching all three translations of *TNR*, and the peptide TSASIGSLCADAR matching the two protein products of *PLP1*.

While the latter two peptides were unambiguous on gene level, the peptide matching the *ENO1* protein product also matched a second IPI entry not associated with the Ensembl gene entry. A further two CSF peptides (RHFFSDYLMGFINSGILK and VATHLSTPQGLQFK) identifying products of *LY6H* and *TNR* respectively, were similar but not identical to HUPO BPP peptides (HFFSDYLMGFINSGILK and VATHLSTPQGLQFKY).

Table 2.16: Signal sequences and transmembrane domains of proteins with peptide evidence in the HUPO BPP and the CSF studies. SS = signal sequence, TM = transmembrane domain, + = present.

Ensembl gene ID	gene symbol	description	SS	TM
ENSG00000130203	*APOE*	Apolipoprotein E precursor	+	
ENSG00000121753	*BAI2*	Brain-specific angiogenesis inhibitor 2 precursor	+	+
ENSG00000132692	*BCAN*	Brevican core protein precursor	+	
ENSG00000162706	*CADM3*	Cell adhesion molecule 3 precursor	+	+
ENSG00000183741	*CBX6*	Chromobox protein homolog 6	+	+
ENSG00000047457	*CP*	Ceruloplasmin precursor	+	
ENSG00000074800	*ENO1*	α-enolase		
ENSG00000111674	*ENO2*	γ-enolase		
ENSG00000108515	*ENO3*	β-enolase		
ENSG00000077782	*FGFR1*	Basic fibroblast growth factor receptor 1 precursor	+	+
ENSG00000131095	*GFAP*	Glial fibrillary acidic protein		
ENSG00000152578	*GRIA4*	Glutamate receptor 4 precursor	+	+
ENSG00000176956	*LY6H*	Lymphocyte antigen 6H precursor	+	
ENSG00000197971	*MBP*	Myelin basic protein		
ENSG00000163531	*NFASC*	Neurofascin precursor	+	+
ENSG00000171246	*NPTX1*	Neuronal pentraxin-1 precursor	+	
ENSG00000091129	*NRCAM*	Neuronal cell adhesion molecule precursor	+	+
ENSG00000126861	*OMG*	Oligodendrocyte-myelin glycoprotein precursor	+	+
ENSG00000184226	*PCDH9*	Protocadherin-9 precursor	+	+
ENSG00000123560	*PLP1*	Myelin proteolipid protein		+
ENSG00000153707	*PTPRD*	Receptor-type tyrosine-protein phosphatase δ precursor	+	+
ENSG00000106278	*PTPRZ1*	Receptor-type tyrosine-protein phosphatase ζ precursor	+	+
ENSG00000116147	*TNR*	Tenascin-R precursor		

A set of six genes common to both studies was identified by peptides specific on gene level. These include *OMG, APOE, BAI2, LY6H, PLP1* and

TNR. While *TNR, PLP1* and *LY6H* were identified by identical or similar peptides, as discussed above, *OMG* and *APOE* had unambiguous peptide matches to different parts of their sequence in the two studies. Unambiguous peptide evidence for *BAI2* was found only in CSF. Table 2.17 summarises the gene-specific evidence for CNS genes common to the CSF and HUPO BPP study. The genes *TNR, LY6H, PLP1* have higher sequence coverage in the HUPO BPP. This could be the result of the collaborative nature of the study, i.e. the higher throughput as well as the diverse range of methodologies applied. Another possible explaination is that the protein products of these genes are less well detectable in CSF then in temporal lobe tissue.

Table 2.17: CNS related genes with specific peptide evidence identified in the HUPO BPP and CSF studies. The listed peptides are specific on gene level.

Ensembl gene ID	gene symbol	description	IPI ID	CSF peptides	HUPO BPP peptides
ENSG00000126861	*OMG*	Oligodendrocyte-myelin glycoprotein precursor	IPI00295832	FTFIPDQSFDQLFQLQEITLYNNR	AHVIGTPCSTQISSLK
ENSG00000130203	*APOE*	Apolipoprotein E precursor	IPI00021842	LKSWFEPLVEDMQR	AKLEEQAQQIR
ENSG00000116147	*TNR*	Tenascin-R precursor	IPI00160552, IPI00514663, IPI00554760	VATHLSTPQGLQFK	VATHLSTPQGLQFKY, DEEEEMMEVSLDATK, LILNYSPR, AEIENYVLTYK, DGQEAAFASYDR, DLMVTASSETSISLIWTK, DSVMVSWSPPVASFDYYR, ELIVDAEDTWIR, GHEFSIPFVEMK, HSQGINWYHWK, ITFTPSSGIASEVTVPK, LDSSVVPNTVTEFTITR, LNPATEYEISLNSVR, LQPPLSQYSVQALRPGSR, MGADGETVVLK,

continued on next page

Table 2.17 – continued from previous page

Ensembl gene ID	gene symbol	description	IPI ID	CSF peptides	HUPO BPP peptides
					NCSEPYCPLGCSSR, SPPTSASVSTVIDGPTQILVR, TSYTLTDLEPGAEYIISVTAER, TTFRLQPPLSQYSVQALRPGSR, VGFGNVEDEFWLGLDNIHR, VSYRPTQVGR, VVYSTLAGEQYHEVLVPR, YEVSVSAVR, YGLVGGEGGR, INFPKK
ENSG00000121753	*BAI2*	Brain-specific angiogenesis inhibitor 2 precursor	IPI00297188, IPI00746807	LLAPAALAFR	-
ENSG00000176956	*LY6H*	Lymphocyte antigen 6H precursor	IPI00014964, IPI00783580	RHFFSDYLMGFINSGILK	HFFSDYLMGFINSGILK, QCQPSDTVCASVR, VDVDCCEK, MCASSCDFVK
ENSG00000123560	*PLP1*	Myelin proteolipid protein	IPI00219661, IPI00396968	TSASIGSLCADAR	MYGVLPWNAFPGK, TSASIGSLCADAR, VCGSNLLSICK

2.4 Discussion

2.4.1 Mapping of protein and peptide identifications

Due to ongoing updates of the genome sequence, changes in genome annotation pipelines, and a steadily increasing amount of sequence information used to define the protein-coding genome, protein sequence space and its representation in databases is constantly changing. It is important to take these changes into account when analysing proteomics experiments in particular if i) data analysis and data generation time are not the same, ii) time separated experiments are to be integrated or iii) experiments differ in the underlying model of the protein space.

In the first case it is important to take changes into account as they affect annotation information associated with proteins identified in a proteomics experiment. Protein annotations - in particular automatic annotations that rely on protein sequence information - are usually based on the current 'version' of the protein sequence space. To analyse a proteomics experiment in context of up-to-data annotations it is therefore necessary to map identifications to the most recent release of a protein database.

In the latter two cases, changes affect the comparability of experiments on the level of the identified protein entries themselves as well as the associated annotations. To compare proteomics studies that have been carried out at different points in time and/or used different releases of the same database or different databases altogether, it is necessary to map experimental data to a common protein sequence space. This ensures consistency in the definition of the identified entities as well as consistent integration with annotations.

In principle, there are three different ways to map protein identifications from one protein sequence space to another. Probably the most commonly used approach is logical mapping using the version history and cross references of a database entry to map between releases and across different databases, respectively. This approach allows propagation between logical entities of different protein sequence spaces which can but do not necessarily have to be, identical with regard to their sequence. Sequence based mapping on the other hand takes the inverse approach to navigate between protein entries with identical sequence that may or may not be the same logical entity. A third alternative available for MS based proteomics data is peptide based mapping to compile a protein set that explains the available peptide evidence. We can either construct a minimal explanatory set by a re-run of the protein identification step using the respective protein identification algorithms, or a maximal explanatory set can be achieved by simple matching of peptide sequences to protein sequences.

Here, protein identifications were mapped between different versions of the same protein sequence space, i.e. different releases of IPI as well as between different protein spaces, i.e. IPI and Ensembl. The mapping served three purposes: i) it allowed integration with the most recent protein annotations, ii) it enabled a more detailed analysis of HUPO BPP peptide evidence with regard to the resolution of brain specific splice isoforms and iii) it allowed comparison between HUPO BPP peptide evidence and peptides identified in an independent study in CSF.

With the respective aims in mind, different approaches were used to map protein identifications. For the analysis of the HUPO BPP dataset in context of biological information, proteins were mapped logically based on IPI entry history to integrate them with current annotations. The emphasis here was on the preservation of the original set of protein identifications. Both the HUPO BPP pilot study, as well as the CSF study, reported minimal explanatory protein sets for the identified peptides that were not immediatly suitable for an in-depth analysis of gene and splice isoform specificity. Thus for this part of the analysis a peptide based mapping approach was taken, resulting in a maximal explanatory set for the observed peptides.

While mapping of proteins based on IPI entry history largely retained the original set of identifications, 5% of identifications were affected by changes in the IPI database. As already discussed, these changes are the result of aforementioned fluctuations in genome annotation and the growing body of sequence information used to construct the IPI database. Both either affect the clustering of sequences in IPI, resulting in the merger of some entries with others, or lead to the deletion of entries from IPI. The latter is caused by entries in the IPI source databases becoming defunct, for example if changes in gene prediction pipelines lead to previously predicted genes being dropped from the genome annotation.

The IPI version used for the re-mapping of peptides to analyse the resolution of CNS specific splice isoforms was separated from the IPI version used for the original identifications by 24 releases, 15 more than the one used for the logic mapping. While the majority of peptides could still be mapped, a fifth of peptide identifications remained "orphan" for reasons already discussed. Interestingly, when the protein set obtained by peptide mapping was compared to that obtained by logical mapping, the logical mapping performed slightly better with 1.1% more proteins in the mapped set. However, the proteins "recovered" by this approach were no longer supported by the original peptide evidence.

This shows that there is no right or wrong approach to move proteomics experiments through time and sequence space. The proteome remains a moving target and peptide as well as protein identification are dependent on the chosen model of protein sequence space at the time of the experiment. However, based on the fact that peptide identifications are closest to the read-out of the experiment in the information flow from MS spectrum to protein identification, it could be argued that the representation of peptide evidence by the current model of sequence space is the more stringent way to move proteomics experiments through time and sequence spaces. This is supported by the observation that some identified IPI entries could be mapped logically to the latest release, but the original peptide identifications no longer matched the mapped entries.

The various proteomics databases take different approaches to address the problem of providing a current view of a proteomics dataset without sacrificing the integrity of the originally reported information. PeptideAtlas, for example uses the protein-coding genome as defined by the Ensembl genome annotation pipeline as a reference. This genome-centric approach maximises stability by firstly having one reference search space, and secondly using a relatively conservative estimate of protein sequence space supported by the rich body of evidence that is used by the Ensembl annotation pipeline. However, the latter also minimises the scope for new findings. Databases like PRIDE, on the other hand, allow the submission of peptide and protein identifications based on any search space. Protein identifications are periodically mapped to the current version of a common sequence space - in this case UniProtKB - based on 100% sequence identity between original protein identification and the mapped sequence entry. The mapped UniProtKB sequence is then used as an entry point for logical cross references to other databases. This approach provides a view of proteomics experiments in context of up-to-date sequence and annotation data whilst at the same time allowing the flexibility of protein identification outside the reference system.

2.4.2 Comparison of protein identifications by experiment, technology and specimen type

The HUPO BPP dataset is characterised by a large variability in the number of proteins identified across the nineteen experiments and five participating laboratories. This consequently translates into a large heterogeneity in protein identifications by different technologies and in different sample types. The majority of proteins reported are single experiment hits. Most of these single-hit identifications originate from two experiments contributed by the same laboratory analysing the same specimen type - biopsy samples - using 1D-PAGE-LC separation.

The clustering results show that the main determinant of "similarity" between protein populations identified in different experiments is sample size rather than any other meta-factor, such as separation technique or specimen type. Samples do not cluster together because of the features (proteins) they have in common. On the contrary, similarity is rather a function of the shared absence of a large number of proteins.

The heterogeneity in the identification number on all levels of the study makes a meaningful comparison of the identified protein sets difficult. Identification sets are complementary rather than confirmatory, and the low re-occurrence of protein identifications across experiments does not allow a statistical interpretation of the performance of the different technologies employed and the different specimen types used with regard to the identifiability of certain proteins or groups of proteins.

It should be noted that - the heterogeneity in size of the reported protein set aside - the experimental design *per se* does not allow a statistical analysis of the results across experiments. An important aspect of good experimental design largely neglected in the study is replication. There are neither technical nor biological replicates within the same lab for any of the technologies and specimen types, which makes it impossible to assess variability. Furthermore, an analysis of reproducibility of protein identifications is hindered by the fact that there is virtually no overlap in technology and specimen types between participating laboratories. Another aspect complicating a comparative interpretation of the study - and also contributing to the heterogeneity in the number of reported protein identifications - is the difference in objectives across experiments. While some experiments were aimed at mapping the protein content of the sample, others focused solely on differentially expressed proteins, consequently narrowing down the set of identifications.

It has to be taken into account, of course, that the number of participants analysing human tissue specimens was small and the number of samples required to achieve an acceptable level of statistical power unrealistic to be produced within the limits of the available resources. However, sacrificing to some extent the diversity of approaches for a more robust experimental design could have increased the benefit of the formidable coordinated effort undertaken by participating laboratories.

2.4.3 Functional composition of the identified protein set

As the heterogeneity of the identified protein sets rendered a meaningful comparative analysis of experiments infeasible, the analysis of the functional composition of the dataset focused on the overall properties of the non-redundant set of proteins identified in the study.

The analysis showed that proteins from a broad spectrum of biological processes and pathways were sampled. It also identified biases towards certain protein classes which partially reflected the tissue origin of the samples and are partially the result of the portfolio of technologies employed in the study.

Consistent with the high vesicle turnover in neurons owing to axonal transport and secretion of neurotransmitters, proteins functioning in vesicle traffic such as motor proteins, components of the vesicle coat as well as proteins involved in vesicle targeting and regulation of vesicle biogenesis and traffic, were abundant. This includes a high number of small GTPases, which probably accounted for a large part of the GTP binding and hydrolase activity observed. Several representatives of the Arf superfamily of GTPases, important regulators of vesicle biogenesis in intracellular traffic [342], as well as members of the Rab GTPases and dynamins, which are involved in vesicle docking [343] and fission respectively, were identified.

Also enriched were proteins involved in transport, specifically in ion and electron transport. To a great extend these activities are related to the enrichment of components of the mitochondrial electron transport chain. Mitochondrial proteins are highly abundant in most cell types in general and in neuronal cell types in particular due to the increased energy demands of brain tissue, accounting for about 20% of the body's oxygen consumption [344].

Identified proteins expressed exclusively in the nervous system, based on available annotations, include the transcriptional repressor scratch 1 (*SCRT1*) and the X-linked transcription factor SOX-3 (*SOX3*), both involved in neuronal differentiation [345, 346]. A third nucleic acid binding protein, Zinc finger protein 312, is uncharacterised. Also nervous system specific is Syntaxin-1B2, a splice isoform of Syntaxin 1. Syntaxin 1 localises to the plasma membrane of axons and synapses [347], where it functions as a t-SNARE, acting in docking and fusion of synaptic vesicles [348].

Due to the selection against hydrophobic and basic proteins inherent to all protein separation techniques employed in the study, transmembrane proteins and nuclear proteins were significantly under-represented in the dataset. This is a problem common to most proteomics studies [349]. However, especially with regard to the study of brain function, both protein classes are of particular importance as they include transcription factors, ion channels and cell surface receptors. The latter also play an important role in drug development, and it should be noted in this context that G-protein coupled receptors, a very important class of drug targets, were almost absent from the dataset.

The strongest bias against transmembrane proteins was observed in experiments that separated the sample by 2D-PAGE prior to identification. However, the results also showed that the application of 1D-PAGE followed by LC, or LC alone also leads - although to a lesser extend - to a significant under-representation of transmembrane proteins.

Representation of the identified IPI entries across the IPI source databases indicated an enrichment of the dataset in relatively well-characterised proteins: two-thirds of all identified proteins had an entry in the manually curated UniProtKB/Swiss-Prot database. This bias is probably rooted in the abundance and biochemical characteristics of these proteins, making them relatively easily accessible to analysis techniques routinely used in protein research. As discussed above, gel-based methods in particular are known to discriminate against certain classes of proteins. However, the results showed that regarding the sensitivity for 'novel' proteins, there was no significant difference in the performance between gel-based and gel-free methods.

That abundance is likely to play a more important role here than biochemical factors is supported by a comparison of the identification frequency of proteins across experiments with mRNA expression levels of the genes encoding them. The results indicate preferential identification of protein products of highly expressed genes.

Proteins identified repeatedly and at high sequence coverage are mainly linked to housekeeping functions such as metabolism, cellular structure and intracellular transport. Identifications with high peptide frequency were clathrin heavy chain (*CLTC*), which is part of the coat of vesicles and coated pits [350], albumin (*ALB*), and heat shock 70kDa protein 12A (*HSPA12A*), which is highly expressed in brain, muscle and kidney. Also frequently identified were several cytoskeleton, cytoskeleton-associated and cytoskeleton-interacting proteins such as microtubule associated protein 2 (*MAP2*); glial fibrillary acidic protein (*GFAP*), a class-III intermediate filament, distinguishing astrocytes from other glial cells [351]; ankyrin-2 (*ANK2*); and neurofilament triplet M protein (*NEF3*).

Proteins identified by all five laboratories included glial fibrillary acidic protein (*GFAP*); α-enolase (*ENO1*), an enzyme functioning in glycolysis; ubiquitin carboxyl-terminal esterase L1 (*UCHL1*), which is involved in the processing of ubiquitin precursors and ubiquinated proteins, and is expressed in the neocortex and the neuroendocrine system; the adaptor protein 14-3-3 ζ (*YWHAZ*); as well as ubiquitously expressed β-actin (*ACTB*).

The biases identified by the functional analysis point out the need for improved sample preparation, fractionation and detection methods that i) do not discriminate against protein classes pivotal to the study of brain function and ii) enhance sensitivity for proteins with low expression levels. In principle there are two approaches to address the latter problem. One is the depletion of highly abundant proteins, a common approach in proteomic analyses of blood plasma samples , which have a particularly large dynamic range. Co-precipitation, however, might result in the removal of proteins potentially important for the understanding of the system under study [352]. Another approach, rapidly gaining importance in current MS based proteomics, is the targeted detection of proteins using protein-specific reporters detected by selected reaction monitoring (SRM) a highly sensitive MS method introduced in the first chapter.

2.4.4 Detection of gene products important for the study of CNS function

Despite the dominance of abundant and well characterised proteins, a detailed assessment of peptide evidence showed that the study successfully identified a significant number of protein products of genes important for the study of CNS function.

The assessment was based on a list of genes considered pivotal to CNS function based on evidence from the literature, curated annotations and expression data. Despite a certain degree of overlap between the different approaches it became clear that information obtained from different sources is currently still mostly complementary rather than mutually supportive. The limited overlap observed here could be explained by the non-exhaustive annotation of UniProtKB/Swiss-Prot entries with regard to tissue specificity, and a lower sensitivity of the literature search and curated annotation concerning CNS specificity of a gene compared to its expression profile. It is to be expected however, that consistency across heterogeneous information sources will improve as curation efforts annotate progressively more published data in databases such as UniProtKB/Swiss-Prot, and as the vast amounts of empirical protein expression data start finding their way into these databases. An example of the latter is given by the recently introduced 'protein existence' tag in the UniProtKB/Swiss-Prot database.

The analysis showed that the HUPO BPP dataset contained expression evidence for a significant proportion of the genes identified as important for CNS function. Many of them have brain specific splice isoforms. Alternative splicing plays an important role in brain physiology and misregulation of splicing is implicated in a number of pathological conditions. Therefore, the detection of splice isoforms is of particular interest with respect to the study of brain function in health and disease.

Resolving splice isoforms, however, is still a challenge in proteomics as inter- and intra-gene sequence similarity due to gene paralogy, conserved domains and isoforms sharing the same exons lead to peptide-to-protein ambiguity. This makes it sometimes difficult to tell which gene was expressed and more commonly which particular isoform of a particular gene was present in the sample.

While in the majority of cases the peptide evidence was specific on gene level, the resolution of splice isoforms was relatively poor. However, for a number of alternatively spliced genes, specific splice isoforms were identified by unambiguous peptide matches, including unambiguously identified brain specific splice isoforms. Some of the unambiguously identified isoforms were, according to UniProtKB/Swiss-Prot annotation, still experimentally unverified on protein level.

Proteomic confirmation of hypothetical and experimentally unverified proteins can greatly improve genome annotation, as it provides experimental evidence on protein level supporting gene predictions and models. The value of mass spectrometry data for the identification, confirmation, and correction of gene annotations has been demonstrated in the model organism *C. elegans* recently [353]. In order for sequence and genome databases, and in turn the proteomics community, to benefit from such information, systematic data collection and integration with sequence databases is an important aspect of large-scale proteomics studies. The aforementioned 'protein existence' tag used in UniProtKB/Swiss-Prot is an example for a mechanism through which proteomics data can feed back to sequence databases. However, the high quality standards applied in Swiss-Prot require thorough quality control of evidence from mass spectrometry experiments, which makes incorporation of proteomic evidence a relatively slow process. To support genome annotation on a large-scale a more efficient feed of proteomics data into the genome annotation process is required. An important component of an infrastructure to effectively interface evidence from proteomics experiments with genome annotation efforts are centralised and publicly available data repositories. Capturing of the vast quantities of identifications and related metadata that are produced by the proteomics community in a systematic and coherent way will greatly facilitate the integration of these data with genome annotation pipelines.

Finally, a comparison between observed isoform specific peptides and predictions of their proteotypic character showed a predictive performance that was no better than random. This result is not entirely surprising as the empirical peptides were not selected based on the prevalence of their identification across the experiments (a hallmark of a true proteotypic peptide), but on their ability to discriminate between proteins (a secondary characteristic of a proteotypic peptide). When the list of empirically observed peptides was restricted to a subsection of frequently identified peptides, the predictions proved to be noticeably albeit not significantly better than random. Interestingly, only about 50% of the empirically frequently observed peptides were in fact predicted to be proteotypic, showing a relatively large false negative rate in the predictions. The results indicate that the fine granularity of predicting the appropriate separation and detection technology for proteotypic peptides did not hold up in the light of the data analysed here.

2.4.5 Conclusions and perspectives

Besides the technological challenges proteomics is facing as an emerging field, another major difficulty it shares with all 'omics' approaches is the extraction of knowledge from the enormous amount of data generated. A comprehensive and consistent analysis of proteomics data in the context of functional annotation adds value to large-scale proteomic studies by putting the results into a biological context. The bird's eye view provided by such an analysis reveals global characteristics of the data not immediately apparent from its parts. This can facilitate evaluation of large-scale proteomics studies, which frequently face challenges in analysing and extracting knowledge from the large quantities of data generated.

While the infrastructure to capture and disseminate experimental data is emerging [256] there is still a lack of bioinformatics resources and tools to support the integrated analysis of proteomic experiments [354] as performed here. As the proteome is still a moving target, the view of a dataset at any given time is only a snapshot based on the current state of protein sequence and annotation resources. Therefore, maintaining an up-to-date view of proteomics data is challenging due to constant change of the underlying resources. However, integration of proteomics data resources with other sources of biomedical knowledge is gradually improving. The increased use of federated data integration facilities like BioMart will facilitate the analysis of large-scale proteomics experiments in the context of biological information from disperse databases. Together with the availability of identifier mapping and cross referencing services [355], they will significantly improve data integration within proteomics as well as across different 'omics' domains.

Besides these independent developments of building blocks for an improved data integration infrastructure there are also ongoing coordinated efforts to connect bioinformatics resources. The Experimental Network for Functional Integration (ENFIN), for example, is a consortium of twenty wet and dry laboratories set up to develop a platform for integration of bioinformatics methods and resources in various domains of biology, and to provide experimentalists with a unified solution to analyse their datasets [356]. Whether a centralised integration strategy as pursued by ENFIN, or federated approaches like BioMart, will ultimatly perform better in the face of the complexity and heterogeneity of the data remains to be seen. More important is the fact that the need for better data integration and analysis tools to support 'omics' research is now widely recognised in the scientific community.

Whith regard to future strategies for an effective study of the human proteome it is apparent that single proteomics experiments currently cannot support the necessary parallelism to achieve the desired amount of data for system biology type of analyses. However, the combined efforts of many laboratories and research consortia worldwide could in fact provide such parallel data generation on a large scale. On the one hand, the complementary nature of the data generated in the HUPO BPP pilot study shows that different analytical technologies can greatly increase the coverage of the system under study. On the other hand however, in light of the study goals, which were the evaluation of different proteomics approaches and specimen types with regard to their suitability to study brain function, the heterogeneity of the reported results raises the question of whether there should have been more emphasis put on experimental design. A relatively loose coordination of experimental approaches applied to the different specimen types, the lack of technical and biological replicates, and the large heterogeneity in the number of protein identifications in the individual experiments, did not in the end, allow the study to address the questions set out to answer. The exploratory nature of the study only allowed to identify overall trends and biases in the dataset which were to a large extent not necessarily specific to the analysis of the brain proteome but common to most current proteomics studies.

"With great power there must also come - great responsibility!" [1] Replicating the success of the human genome project in proteomics will require a high degree of collaboration and coordination to overcome the analytical challenges involved in defining the proteome. To funnel the power and maximise the synergistic effect of large collaborative endeavours like the HUPO BPP, the proteomics community will have to trade in, at least to some extent, academic freedom for clearly defined targets and a more rigorous study design. Only then can a transition from exploratory to confirmatory data analysis be achieved allowing collaborative studies to become more than the sum of their parts.

[1] Benjamin Parker in Amazing Fantasy #15 (August 1962) - The first Spider-Man story

Chapter 3

Proteomic footprint on the genome: a survey of expression evidence for protein-coding genes in public proteomics repositories

3.1 Introduction

Delineating the protein-coding genome is central to the understanding of human biology on the molecular level. Functional diversification on protein level arising from alternative transcription [357, 358] and translation [16] as well as post-translational modification of protein products makes the study of gene function on protein level indispensable.

Given the potential extent of combinatorial diversity, together with the variation of protein isoforms and modifications across tissues and physiological states [359, 360], the task of mapping the proteome is immense.

A first step towards a comprehensive annotation of the proteome is the identification of protein-coding genes and the protein products they encode. The recent announcement by the UniProt consortium of the completion of the first draft of a fully curated representation of the complete human protein-coding genome in Swiss-Prot [361] marked a milestone on the path to achieving this goal. This manually reviewed set of currently known protein-coding genes consists of 20332 entries covering 34129 distinct protein sequences if annotated splice isoforms are included (release 14.7, January 2009).

Most protein sequences in the UniProt Knowledgebase (UniProtKB) are derived from the translation of coding sequences submitted to one of the three nucleotide sequence databases European Molecular Biology Laboratory (EMBL)-Bank, GenBank or DNA Data Bank of Japan (DDBJ). The strength of evidence for the existence of a protein product differs from gene to gene, and a recently introduced 'protein existence' attribute indicates the degree of confidence [362]. While some entries are based purely on gene predictions, the existence of others is inferred from homology to known proteins in closely related species. For the majority of human Swiss-Prot entries reviewed experimental evidence exists either on transcript (37.1%) or directly on protein level (58.9%) in the form of sequence information from Edman degradation, X-ray or nuclear magnetic resonance (NMR) structure, antibody detection or identification by mass spectrometry (MS).

In particular, in light of its high-throughput capabilities, the latter is an important technology to assist the annotation of protein sequences on a large-scale. Evidence from mass spectrometry experiments can not only confirm that a gene is translated into a protein sequence. Peptides identified by MS can also be used to confirm gene models [147] as well as to locate sites of post-translational modification [154, 180]. Mass spectrometry has also been successfully employed to identify previously unknown genes by matching mass spectra against the six frame translations of genomes [150, 363].

Systematic integration of MS proteomics data with the genome sequence, however, is still sparse and its potential as supporting evidence in the genome annotation process practically untapped [151]. Although the use of mass spectrometry data as a tool for genome annotation has been demonstrated [364, 365, 148] its integration with the genome sequence remains largely restricted to retrospective mapping of peptide identifications to annotated genomes. The 'protein existence' tag in UniProtKB entries is one example of proteomics data being used as confirmatory evidence for existing protein records [362]. Proteomics resources like PeptideAtlas [366, 257] and the Global Proteome Machine Database (gpmDB) [257] map or base their peptide identifications on sequences translated from existing gene models of the Ensembl database. Peptide identifications from the Max-Planck Unified Proteome Database (MAPU) are integrated with Ensembl protein entries via the Distributed Annotation System (DAS) [367].

Like the efforts to integrate proteomics data with the genome, proteomics data itself is disperse. Although several proteomics repositories have agreed to exchange information between their databases in the future [267], currently there is no integrated view of proteomics data in the public domain and none of the major genome browsers offers systematic integration of proteomics information on the genome sequence as it is available for transcript or other types of 'omics' data already [9, 368, 369, 370, 371, 372], not least because these data are systematically used in genome annotation.

Meanwhile, plans for a Human Proteome Project (HPP) to systematically map the proteome have emerged [373]. The drafting process of the project is accompanied by discussions of whether a HPP should be gene- or protein-centric [374] and it is argued that the suitability of mass spectrometry for the former would be limited [241].

To assess the potential, as well as the limitations, of current MS based proteomics as a tool for genome annotation, a survey of proteomics data in the public domain was conducted. The aim of the analysis was to get a clear picture of i) the combined status of proteomics efforts to date with regard to the coverage and resolution of the human protein-coding genome and ii) the extent of bias and redundancy found in proteomics datasets discussed in the previous chapter, on a global level.

Peptide identifications were obtained from the four proteomics repositories PeptideAtlas [149], gpmDB [257], Human Proteinpedia (HuPA) [375] and the Protein Identification Database (PRIDE) [256]. The four resources differ in their goals and philosophy. PeptideAtlas and gpmDB aim to build libraries of proteotypic peptides [376] and their fragmentation patterns by applying a uniform analysis process to identify peptides from MS spectra submitted to their databases. Both resources integrate MS data with a single reference genome, PeptideAtlas, by mapping peptide identifications to Ensembl protein entries [366] and gpmDB by using Ensembl protein sequence information for peptide identification [257]. HuPA and PRIDE are repositories for MS proteomics experiments accepting submission of peptide and protein identifications made by the data submitter. Thus, identifications are based on a range of different algorithms and databases. PRIDE maps peptide and protein identifications periodically to a common sequence space for integration with up-to-date sequence and annotation information. However, the peptide information available for download is provided unprocessed as deposited by the original data submitter.

Peptide data obtained from the four databases was integrated by aligning it to the protein-coding genome as defined by the Ensembl genome annotation. Using Ensembl as a reference space for integration put the proteomics data in a single genomic context with the added benefit of transitive integration with the large body of genome information and annotations available through the Ensembl project including gene models, single nucleotide polymorphisms (SNPs), microarray probeset mappings as well as functional annotations. This enabled a straightforward analysis of peptide data regarding qualitative and quantitative coverage of the genome, i.e the proportion of genes and proteins covered by proteomics evidence, confirmatory evidence for gene models on exon and splice site level, detection of coding SNPs (cSNPs), redundancy of peptide identification and biases in the sampling of functional categories. The results of this analysis are presented and discussed in this chapter.

3.2 Materials & methods

3.2.1 Peptide sequences

Peptide data was obtained from the four MS proteomics resources PeptideAtlas, gpmDB, HuPA and PRIDE. Data from PeptideAtlas, gpmDB and HuPA was downloaded as flat text files from the respective hypertext transfer protocol (HTTP) or file transfer protocol (FTP) servers. Peptide identifications from PRIDE were obtained via the PRIDE BioMart. See table 3.1 for release information and universal resource locators (URLs).

Table 3.1: Origin of peptide data used in the survey. The table lists the database name the database release (PeptideAtlas, HuPA) or downloaded date (gpmDB, PRIDE), and the URL the data was obtained from.

database	release/date	URL
PeptideAtlas	April 2007	http://www.peptideatlas.org/builds/human/HumanP0.9/200704/APD_Hs_all.fasta
gpmDB	05/09/2008	ftp://ftp.thegpm.org/proteotypic_peptide_profiles/eukaryotes/GPMp/human_cmp_20.fasta
HuPA	2.0	http://www.humanproteinpedia.org/HuPA_Download/FULL/HUPA_RELEASE_2.0.zip
PRIDE	03/09/2008	http://www.ebi.ac.uk/pride/prideMart.do

From the HuPA database only cell line and tissue expression data contained in the files HUPA_Cell_line_Expression_MS_PEPTIDES.txt and HUPA_Tissue_Expression_MS_PEPTIDES.txt was used in the analysis.

For use with the basic local alignment search tool (BLAST) (see section 3.2.4) peptide sequences obtained from HuPA and PRIDE were converted to FASTA format [377].

3.2.2 Protein sequences, single nucleotide polymorphisms and exons

Protein entries of Ensembl (release 50) were obtained in FASTA format from the Ensembl FTP site (ftp://ftp.ensembl.org/pub/release-50/fasta/homo_sapiens/pep/Homo_sapiens.NCBI36.50.pep.all.fa.gz). A non-redundant (nr) protein sequence file in FASTA format was created by grouping protein entries with identical amino acid sequence resulting from translation of Ensembl transcripts with identical coding sequence (CDS).

Protein sequence coordinates of non-synonymous coding single nucleotide polymorphism (cSNP)s were obtained from Ensembl by querying the Ensembl Mart schema directly via the public MySQL database instance (martdb.ensembl.org:5316/ensembl_mart_50).

Protein sequence coordinates of exon boundaries were obtained by querying Ensembl through the Ensembl Perl application programming interface (API) (http://www.ensembl.org/info/docs/api/index.html).

3.2.3 Transcript expression data

gcRMA normalised expression data [333] was downloaded from the SymAtlas [229] website (http://wombat.gnf.org/downloads/gnf1h, gcrma.zip). Expression values were mapped to Ensembl genes using the mapping of Affymetrix probeset identifiers to Ensembl genes provided by the Ensembl database or in the case of GNF custom probesets based on the Genomics Institute of the Novartis Research Foundation (GNF) chip annotation tables (http://wombat.gnf.org/downloads/gnf1h-anntable.zip).

3.2.4 Sequence alignment

Peptide sequences were aligned to Ensembl protein sequences using the local sequence alignment program BLAST [262].

BLAST executables (version 2.2.18) were downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/blast-2.2.18-ia32-linux.tar.gz.

A non-redundant FASTA file of Ensembl protein sequences (see section 3.2.2) was used as the search database. Amino acid residues affected by non-synonymous cSNPs were masked by replacing them with the one-letter amino acid code 'X' representing an unknown amino acid.

Parameters optimised to find matches to short peptides (http://www.ncbi.nlm.nih.gov/blast/producttable.shtml#shortp) were used for the BLAST search:

1. `word size = 2`
2. `SEG Filter = Off`
3. `expectation value = 20000`
4. `score matrix = PAM30`

The results were subsequently filtered for ungapped alignments stretching the entire query peptide sequences. Mismatches were only allowed at positions of non-synonymous cSNPs and the mismatching peptide residue had to be an allowed single nucleotide polymorphism (SNP) allele.

3.2.5 *In silico* proteolytic digest

Theoretical tryptic peptides were generated using the proteomics software toolkit DBToolkit (version 3.5.4) [378]. A FASTA file of non-redundant Ensembl protein sequences (see section 3.2.2) was used as input. Missed cleavages were not allowed and peptide sequences were filtered by theoretical molecular mass retaining only peptides in the mass range between 600 - 4000 Da.

3.2.6 Gene Ontology analysis

The Gene Ontology analysis was performed as described in section 2.2.7.

3.3 Results

3.3.1 Peptide data obtained from PeptideAtlas, gpmDB, HuPA and PRIDE

Peptide sequence information was obtained from the four MS proteomics data bases PeptideAtlas, gpmDB, HuPA and PRIDE. Table 3.2 summarises the number of peptide records and non-redundant peptide sequences retrieved from each database and in total.

As the four resources differ in their goals and philosophy, the peptide information provided by the four databases is different in nature. PeptideAtlas and gpmDB provide identified peptide sequences as non-redundant sequence libraries of proteotypic peptides, i.e. peptides that are frequently and repeatedly observed for a given protein. Information about the identification frequency of individual peptides is not supplied with the sequences.

Peptide information from PRIDE and HuPA was obtained as redundant lists of peptide sequence records. The frequencies of individual peptide sequences in the list are equivalent to the number of times a particular peptide has been reported as identified in the database. On average 116000 nr peptide sequences were obtained from each database. The smaller number of peptide sequences contained in PeptideAtlas can probably be explained by the most recent build available at the time of analysis being two years old.

Table 3.2: Peptide sequence data obtained from the four MS proteomics repositories PeptideAtlas, gpmDB, HuPA an PRIDE.

database	peptide records	peptide sequences (nr)
PeptideAtlas	84086	84086
gpmDB	124655	124655
HuPA	1668736	115634
PRIDE	1598676	140016
total	3476153	280703

The total number of non-redundant peptide sequences, shown in table 3.2, indicated that peptide data available from the four resources was to a large extent complementary. Table 3.3 summarises the overlap of sequence information between databases. The average overlap was ∼24%. Only 7% of all peptide sequences were found in all four databases and 57% were unique to the respective resource. With an overlap of 44%, PRIDE and HuPA were the two databases with the highest number of shared peptide information. The lower overlap of PeptideAtlas and gpmDB with HuPA and PRIDE respectively, could be explained by the different nature of peptide information provided by the first two databases limiting sequence information to that of proteotypic peptides as defined by a single processing pipeline.

Table 3.3: Overlap of peptide sequence information obtained from PeptideAtlas, gpmDB, HuPA and PRIDE. The overlap between two databases is defined as the number of peptides found in the intersection of the peptide sets from the respective databases and is given as the percentage of the union of the peptide sets. Unique peptides are defined as peptides which are unique amongst all nr peptides found in the four databases. For the individual databases the proportion of unique peptides is given as the percentage of all nr peptide sequences in the database. The total proportion of unique sequences is the sum of all unique peptides given as the percentage of all nr peptide sequences in the four databases.

database	HuPA		PeptideAtlas		PRIDE		unique	
gpmDB	36683	(18%)	43272	(26%)	44391	(20%)	60016	(48%)
HuPA			29353	(17%)	79055	(44%)	27164	(23%)
PeptideAtlas					34140	(18%)	30701	(36%)
PRIDE							43315	(31%)
overlap between all	18974	(7%)			total unique		161196	(57%)

Information about the proteases used to generate the obtained peptides was either not available or not obtained from the respective databases. However, the protease of choice in the majority of proteomics experiments is trypsin due to its high sequence specificity and efficiency. Furthermore, in tryptic peptides basic amino acid residues reside at the carboxy (C)-terminus, wich is beneficial for sequencing by collision induced dissociation (CID). The distribution of the C-terminal residue frequency (figure 3.1) shows that the vast majority of peptides (84%) indeed end in the residues arginine (R) or lysine (K) and are thus likely to result from proteolysis with trypsin or the less frequently used proteases Arg-C and Lys-C. Other proteases that find use in MS based proteomics experiments are pepsin A (cleaves C-terminal of F, L), V8E (cleaves C-terminal of E) and V8DE (cleaves C-terminal of E, D). However, given the infrequent use of these proteases, it is more likely that the majority of peptides with non-tryptic ends are either peptides originating from the C-terminal end of digested proteins or truncations of tryptic peptides.

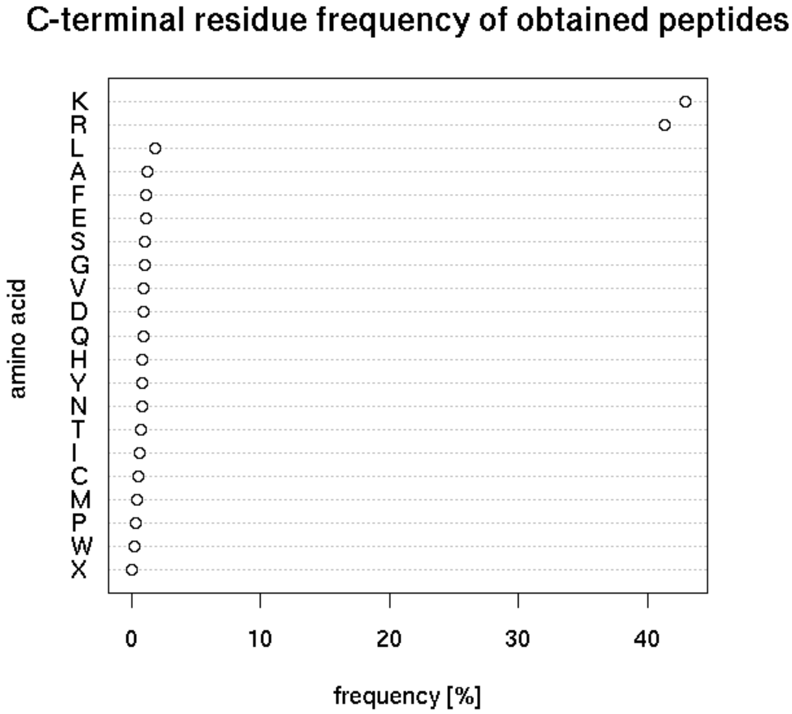


Figure 3.1: Frequency distribution of carboxy-terminal peptide residues. Residues are specified by their one-letter amino acid code.

3.3.2 Integration of peptide data with the genome

3.3.2.1 Alignment of peptides to Ensembl protein sequences

Peptide sequences were put into genomic context by aligning them to the non-redundant set of translations of Ensembl protein-coding genes. Residues in the protein sequence affected by cSNPs were masked to allow straightforward alignment of peptide sequences with different SNP alleles. Alignments were subsequently filtered for perfect matches. Residues matching at masked SNP positions had to be an allowed consequence of one of the nucleotide SNP variants.

The results of the peptide sequence alignment are summarised in table 3.4. The ∼215000 (76.8%) successfully aligned sequences originated from 2.8 million peptide records, which was 81% of all records obtained. Unaligned sequences were most likely the result of sequence changes over time as well as differences between the databases used for peptide identification. These occurances are discussed in detail in the previous chapter.

The ∼10% higher alignment success rate for peptide sequences from PeptideAtlas and gpmDB is probably explained by the fact that both databases use the Ensembl genome as a reference. Failed alignments could be explained by sequence changes between Ensembl versions as well as the fact that both databases also contain peptide identifications based on other sequence databases like the International Protein Index (IPI) and NCBInr [275] that are not always mappable to Ensembl protein entries.

Table 3.4: Peptides aligned to Ensembl protein sequences.

database	aligned peptide records	aligned peptide sequences (nr)
PeptideAtlas	69922 (83.2%)	69922 (83.2%)
gpmDB	106103 (85.1%)	106103 (85.1%)
HuPA	1339524 (80.3%)	85856 (74.2%)
PRIDE	1301326 (81.4%)	104127 (74.4%)
total	2816875 (81.0%)	215552 (76.8%)

3.3.2.2 Properties of aligned peptide sequences

A sequence match *per se* does not mean that the respective peptide sequence originates from the matched stretch of protein sequence. Whether or not a peptide can theoretically originate from the matched location depends on the specificity of the protease and thus the context of the sequence match. To be a candidate origin for a matching peptide sequence, a match location has to be flanked, or in the case of N-terminal and C-terminal fragments, followed or preceded by valid protease recognition sites. In cases where these conditions are not met there is also the possibility of the peptide sequence arising from break-up of a proteolytic peptide.

As protease information was not obtained with the peptide sequences the proteolytic character of matching peptides had to be empirically determined from the data. Successfully aligned peptides showed a similar frequency distribution of C-terminal residues as the overall distribution observed for all peptides obtained from the four databases (figure 3.1) with a prevalence of tryptic cleavage sites (figure 3.2 A). Because of the very low frequency of residues implicating the use of other proteases only trypsin was taken into account as a peptide generating protease.

Previous analyses of trypsin specificity suggested that the protease cleaves exclusively C-terminal of lysine and arginine. The only non-tryptic peptides likely to be observed are break-up products of fully tryptic peptides N-terminal to an internal proline [54]. Therefore, peptide matches were filtered further for peptide sequences originating from fully tryptic locations (flanked by arginine or lysine residues which are not followed by proline) or C-terminal fragments of such peptides resulting from truncation occuring N-terminal of internal prolines that are preceded by aspartic acid (the amide bond between D-P residues in peptides is the weakest and thus easily hydrolisable [54]).

As figure 3.2 B shows the majority of alignabed peptides ($\sim$69%) originated from fully tryptic peptide locations with only a small minority of 459 peptides (0.2% of all aligned peptides and thus not visible in the bar plot) potentially resulting from truncation of tryptic peptides. A similar relative

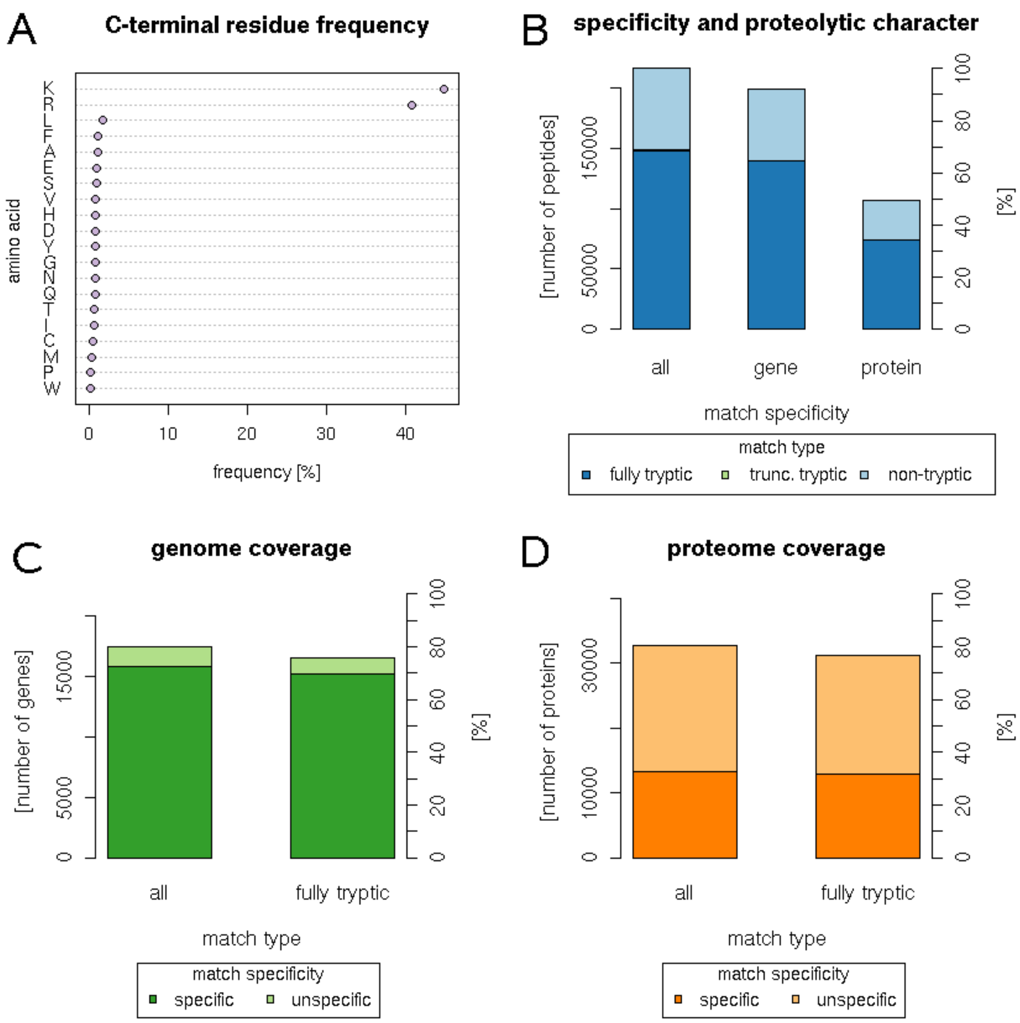


Figure 3.2: Properties of aligned peptide sequences. A) C-terminal residue frequency. B) Specificity and proteolytic character. Fully tryptic peptides are peptides originating from at least one protein sequence location flanked by K or R including those with internal K or R residues. Truncated (trunc.) tryptic peptides are C-terminal fragments of fully tryptic peptides that have been truncated N-terminal of proline residues preceeded by aspartic acid. C) and D) Gene and protein coverage. The proportion of genes/proteins that have at least one peptide match is shown for all peptides and for fully tryptic peptides. Specific peptide matches uniquely identify a given gene or protein.

frequency of tryptic peptides was observed for gene and protein-specific peptides, which make up $\sim$90% and $\sim$50% of all aligned peptides, respectively. Here, peptide matches are defined as *protein-specific* if they match only a single protein sequence and as *gene-specific* only if they match protein sequences encoded by the same gene. Tryptic peptide matches are defined as protein-specific if they match a fully tryptic sequence location in exactly one protein sequence. They are considered gene-specific if they match a fully tryptic sequence location only in protein products originating from exactly one gene.

Figures 3.2 C and D show the proportions of protein-coding genes and protein products matched by all aligned as well as fully tryptic peptides only. Around 80% of genes and proteins were "touched" by peptide evidence, i.e. had at least one peptide match. Peptides matching only products of a specific gene were present for $\sim$72% of all protein-coding genes while peptides uniquely matching a single protein product were present for $\sim$32% of all protein sequences. As the bar plot shows, considering fully tryptic peptides alone only marginally affected the overall coverage on both gene and protein level.

The above results show that tryptic peptides were prevalent, and that considering only fully tryptic peptides did not significantly affect gene and protein coverage. This together with the results of previous experimental studies, suggesting that for a typical shotgun proteomics experiment only fully tryptic peptides should be used for peptide identification [54], were the rationale to consider only fully tryptic peptide matches for further analysis.

3.3.2.3 Properties of fully tryptic peptide sequences

Although trypsin is a highly specific protease, tryptic digests are never 100% efficient, resulting in partially cleaved fragments with internal cleavage sites. Figure 3.3 A shows the frequency distribution of missed cleavage sites for observed fully tryptic peptides. Half of all fully tryptic peptides were perfect tryptic peptides, i.e. did not have internal missed cleavage sites. Almost 98% of all peptides had a maximum of two missed cleavage sites. The steep drop in the frequency of peptides with more than two internal missed cleavage sites is consistent with most search engines allowing for a maximum of two missed cleavages. This is to reduce the number of random matches resulting from the higher number of theoretical peptide masses to be matched against the experimental data.

An *in silico* tryptic digest of Ensembl protein sequences, not allowing for any missed cleavages to simulate a perfect proteolysis, resulted in 622709 distinct peptide sequences in the mass range between 600 - 4000 Dalton (Da). As figure 3.3 B shows 20% of theoretical peptides have been observed either directly as perfect tryptic peptides, indirectly as subsequences of larger tryptic peptides resulting from missed cleavage, or both directly and indirectly.

Figure 3.3 C shows the peptide mass distribution of all theoretical perfect tryptic peptides in comparison to theoretical peptides covered by experimental peptides and actually observed fully tryptic peptides. The distribution of peptide masses of observed tryptic peptides was skewed towards higher masses, and slightly more evenly spread, with a less pronounced peak at lower peptide masses. The mass distribution of theoretical peptides covered by observed peptides indicates that the shift towards higher peptide masses is not only the result of missed cleavages but an overall shift of the distribution towards higher masses. Stochastically, peptides with higher masses are more likely to be identified as the frequency of background peptides with similar masses is lower at higher mass ranges. Secondly, identifications of longer peptide sequences are statistically more significant.

Figure 3.3 D shows the degeneracy of all theoretical perfect tryptic peptides in comparison to experimentally observed fully tryptic peptides. The distribution of experimentally observed peptides across the different levels of degeneracy is very similar to that of theoretical tryptic peptides with around 50% of observed peptides being non-degenerate.

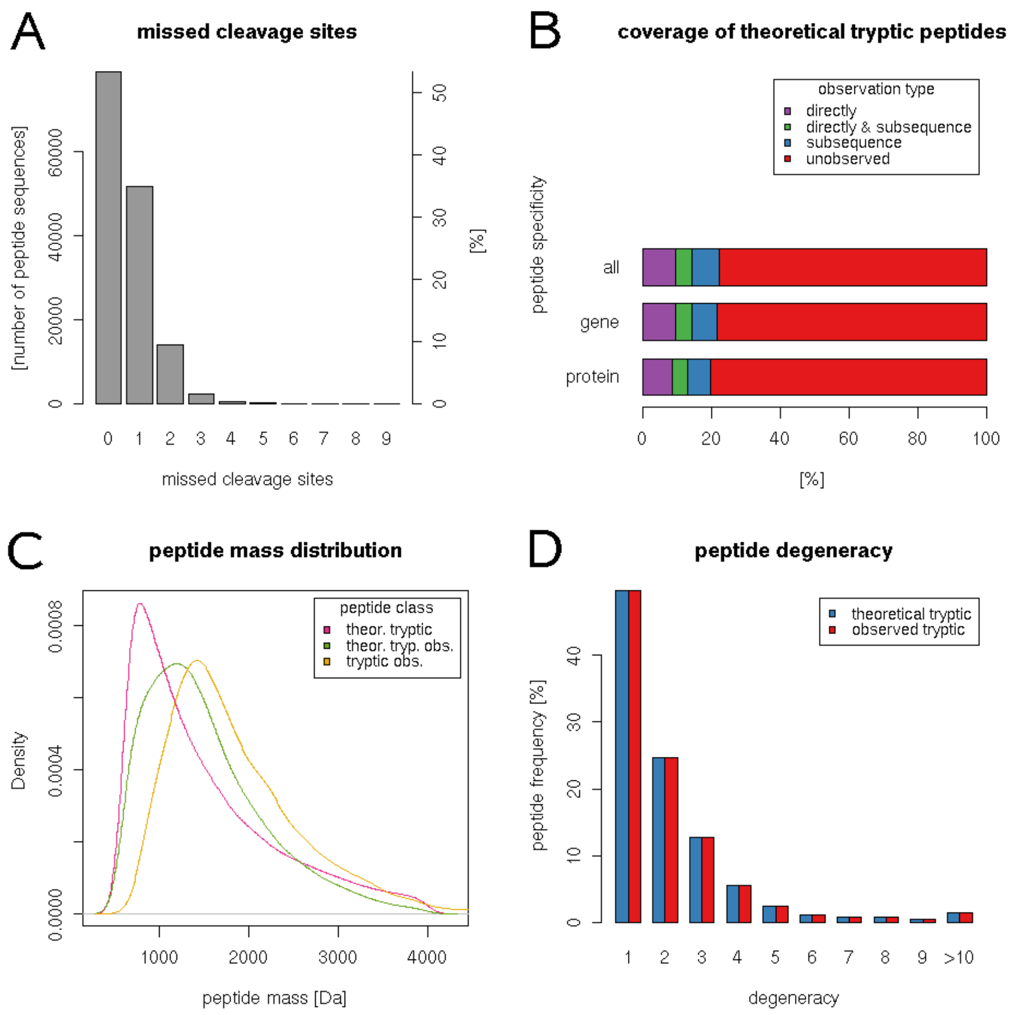


Figure 3.3: Properties of aligned fully tryptic peptide sequences. A) Frequency of missed cleavage sites in fully tryptic peptides. Shown is the frequency distribution of missed cleavage sites for fully tryptic peptide sequences. B) Coverage of theoretical tryptic peptides. Shown is the proportion of theoretical perfect, fully tryptic peptides without missed cleavages covered by experimentally observed peptides. The theoretical peptide sequence could either be observed directly or as a subsequence of an observed peptide with one or more missed cleavages, or both. C) Probablity density function of the masses of i) theoretical tryptic peptides, ii) theoretical tryptic peptide observed directly or as a subsequence, and iii) the experimentally observed tryptic peptides. D) Degeneracy of experimentally observed tryptic peptides *vs* theoretical tryptic peptides. The degeneracy of a peptide is given by the number of protein sequences it maps to (x-axis). The bar height shows the frequency of peptide sequences at each level of degeneracy.

3.3.3 Genome coverage of fully tryptic peptides

Next, coverage of the genome by tryptic peptide evidence was analysed in more detail looking at confirmatory evidence for gene models and cSNPs, identification frequencies of individual genes and functional classes of genes as well as evidence overlap between databases with regard to genome coverage.

3.3.3.1 Confirmatory evidence for gene models

Figure 3.4 summarises confirmatory evidence provided by tryptic peptides for genome annotation on gene, protein, exon and splice site level.

Tryptic peptides covered 76% of the 21785 protein-coding genes annotated in Ensembl. For 69% of all genes peptide evidence was gene-specific. 44% of all protein-coding genes encode alternative protein products. Gene-specific peptide evidence covered 81% of these genes while 59% of unspliced genes had one or more specific tryptic peptide matches. This left 24% of the protein-coding genome “untouched” by tryptic peptide identifications.

On protein level the overall coverage was similar with 77% of the 40699 distinct protein sequences “touched” by tryptic peptide evidence. However, only a third of proteins were identified by protein-specific peptides matching exactly one sequence. As the pie chart shows, the majority of proteins covered only by unspecific peptide evidence were protein products of alternatively spliced genes. Intra-gene sequence similarity between splice isoforms renders a large number of peptides emitted by these proteins ambiguous, leaving only 30% of all theoretical tryptic peptides originating from alternative protein products for unambiguous protein identifications. Yet, 17% of the 29001 proteins translated from alternatively spliced messenger RNA (mRNA)s were identified by such isoform specific peptides. Of the 11764 proteins encoded by unspliced genes 62% are identified by protein-specific peptides.

Drilling one level deeper to the 235626 exons making up the translated

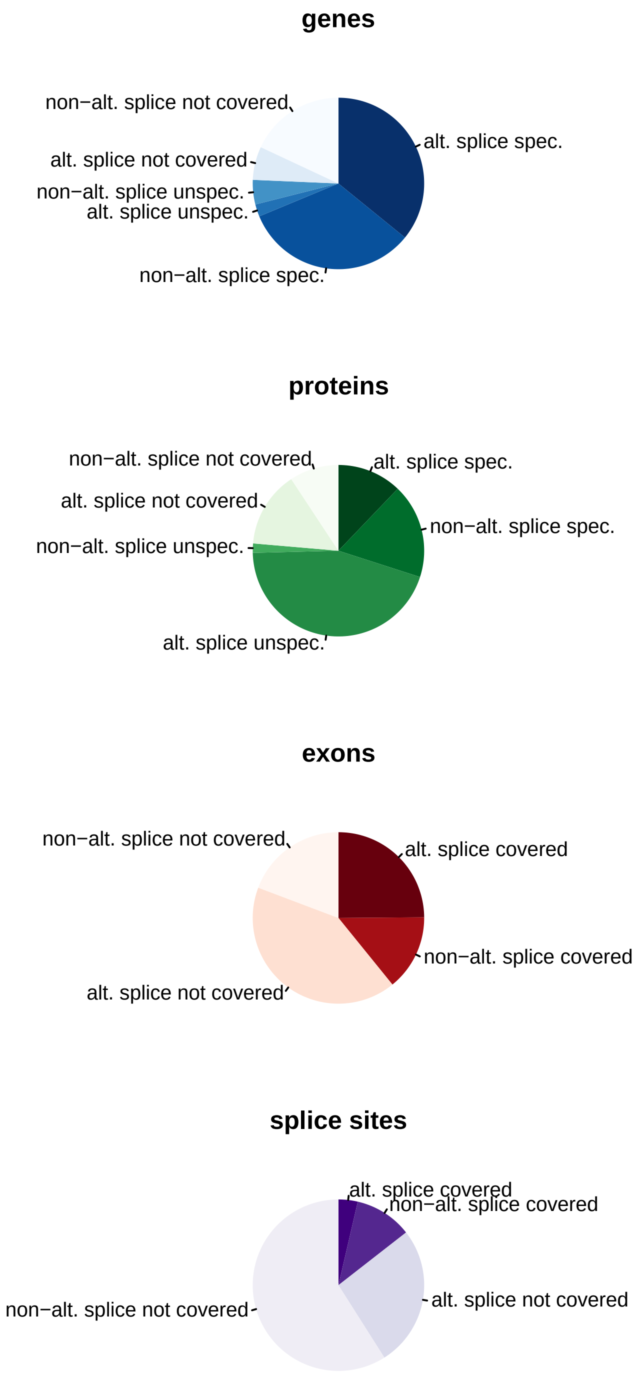


Figure 3.4: Gene, protein, exon and splice site coverage of fully trypic peptides. alt. = alternatively, spec. = specific.

regions of protein-coding mRNA, tryptic peptides provided evidence for the usage of 92270 (39%) exons. That is 37% of all exons of alternatively spliced and 43% of all exons of non-alternatively spliced genes.

Peptides that coincided with the location of exon boundaries in the protein sequence provided evidence for splice events. Experimentally observed tryptic peptides bridged 14% of the 222169 splice events occuring between coding exons; 15% of the splice events covered were alternative events (where the splice donor site can be spliced with more than one acceptor site) and 85% non-alternative events (i.e. splicing always occurs between the same two exons).

Figure 3.5 provides a more detailed summary of gene model confirmation by tryptic peptide evidence on exon and splice event level. Around 2400 genes had more than 95 % of their exons confirmed by tryptic peptide matches (figure 3.5 A) and 2384 genes (11% of all protein-coding genes) - 251 alternatively spliced and 2133 unspliced genes - had a peptide match to all of their exons. The histogram in figure 3.5 B shows, however, that the majority of these genes were genes with a single coding exon only. The gene with the largest number of coding exons confirmed by tryptic peptide evidence was the serine/threonine kinase *ATR*, with 47 exons.

On splice event level confirmatory evidence from MS was much sparser, since the frequency of peptides coinciding with exon boundaries, and thus the chance of observing them is lower (figure 3.5 C). Nevertheless, 200 genes were matched by peptides that provided protein level evidence for all splice events occuring between their exons. A small proportion of ten genes were alternatively spliced. The remaining 188 genes encoded only one protein product. Similar to the observation on exon level, the majority of genes with full coverage here also had relatively simple gene models with only one or two splice events (figure 3.5 D). The Glutamate dehydrogenase 1 (*GLUD1*) gene, which encodes two protein isoforms, was with thirteen covered splice events the gene with the highest number of fully covered splice events. The gene with the most complex gene model confirmed by peptides covering six alternative splice events was the Histone

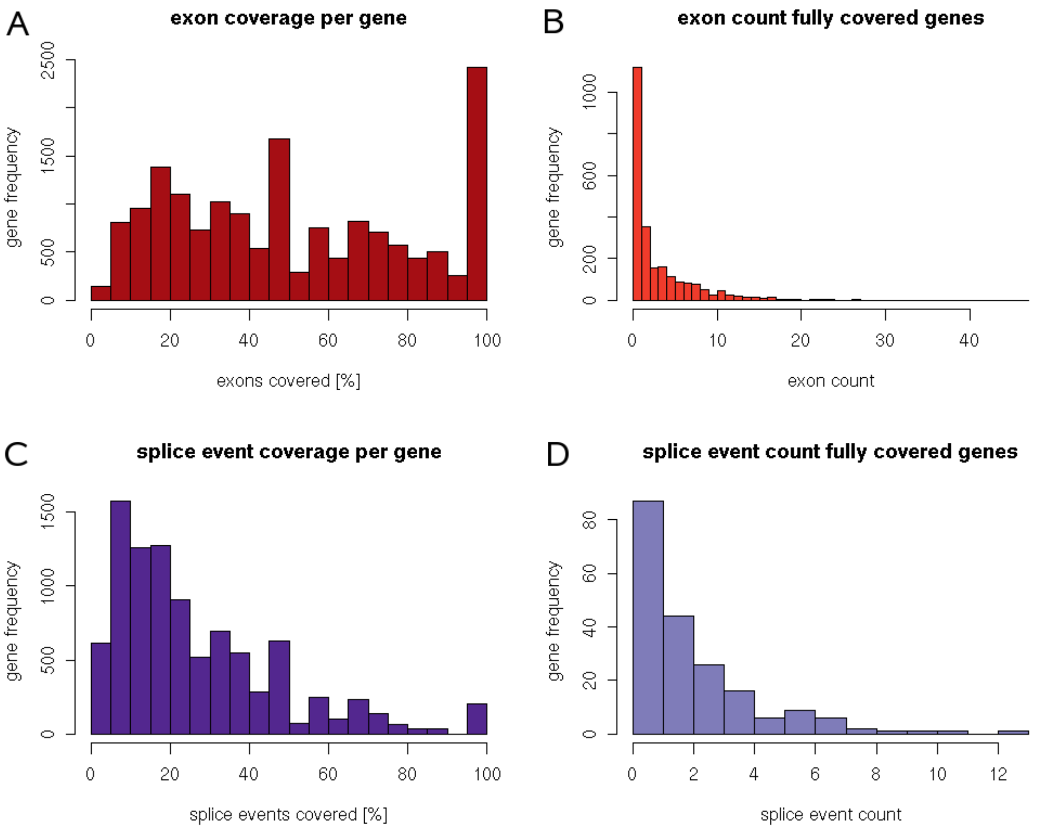


Figure 3.5: Gene model confirmation on exon and splice site level by peptide evidence in the PRIDE and HuPA databases. A) Histogram of the proportion of exons covered per gene. B) Histogram of the exon count of genes with peptide evidence covering all their exons. C) Histogram of the proportion of splice events covered per gene. D) Histogram of the splice event count of genes with peptide evidence covering all their splice events.

gene *H3F3A* which encodes three different protein isoforms. Table 3.5 lists the genes encoding alternative protein isoforms that had all splice events confirmed by tryptic peptides.

3.3.3.2 Evidence overlap between databases

Tables 3.6 and 3.7 summarise the overlap of peptide information between PeptideAtlas, gpmDB, PRIDE and HuPA at the level of observed tryptic peptide sequences and theoretical tryptic peptides (directly or indirectly)

Table 3.5: Gene models of alternatively spliced genes confirmed by tryptic peptides. alt. = alternative, non-alt. = non-alternative.

gene symbol	description	Ensembl ID	iso-forms	alt. splice events	non-alt. splice events
H3F3A	Histone H3.3	ENSG00000163041	4	6	0
PVALB	Parvalbumin α	ENSG00000100362	3	7	0
ISG20	Interferon-stimulated gene 20 kDa protein	ENSG00000172183	3	3	1
GLUD1	Glutamate dehydrogenase 1, mitochondrial Precursor	ENSG00000148672	2	2	11
AK3	GTP:AMP phosphotransferase mitochondrial	ENSG00000147853	2	4	2
PHB	Prohibitin	ENSG00000167085	2	2	4
UBE2G1	Ubiquitin-conjugating enzyme E2 G1	ENSG00000132388	2	2	3
FABP1	Fatty acid-binding protein, liver	ENSG00000163586	2	4	1
FABP5	Fatty acid-binding protein, epidermal	ENSG00000164687	2	4	1
KNG1	Kininogen-1 Precursor	ENSG00000113889	2	2	9

covered by observed peptides. Compared to the overlap of unfiltered sequence information obtained from the four databases (see table 3.3) the pairwise as well as overall overlap was higher on tryptic peptide level. The average overlap between individual databases was 31%. Peptides found in all four databases amounted to 11% of the 147947 fully tryptic peptides while half of these peptides were reported in a single database only.

Agreement between databases improved further if the comparison was made on the level of theoretical tryptic peptides covered by observed peptide sequences. Average pairwise overlap went up to 33% and overall overlap between all four databases up to 12%. The proportion of peptides covered by one database only went down to 44%.

Figures 3.6 and 3.7 show the overlap in gene and protein coverage of fully tryptic peptide evidence from PeptideAtlas, gpmDB, HuPA and PRIDE. If genes were excluded that do not emit gene-specific tryptic peptides (grey

Table 3.6: Overlap of fully tryptic peptide sequence information obtained from PeptideAtlas, gpmDB, HuPA and PRIDE. See legend of table 3.3 for explanation of how percentages were calculated.

database	HuPA		PeptideAtlas		PRIDE		unique	
gpmDB	28834	(25%)	29811	(30%)	33093	(24%)	29593	(36%)
HuPA			21510	(24%)	57049	(59%)	1979	(3%)
PeptideAtlas					25086	(22%)	13986	(28%)
PRIDE							23004	(26%)
overlap between all	15816	(11%)			total unique		68562	(50%)

Table 3.7: Overlap of theoretical tryptic peptides covered by sequence information obtained from PeptideAtlas, gpmDB, HuPA and PRIDE. See legend of table 3.3 for explanation of how percentages were calculated.

database	HuPA		PeptideAtlas		PRIDE		unique	
gpmDB	31231	(30%)	29740	(35%)	35266	(29%)	22036	(32%)
HuPA			23048	(26%)	54891	(55%)	5778	(9%)
PeptideAtlas					26272	(24%)	10817	(23%)
PRIDE							22322	(25%)
overlap between all	17118	(12%)			total unique		60953	(44%)

shaded area) due to inter-gene sequence similarity, unambiguous peptide evidence was reported for 73% of the protein-coding genome in at least one proteomics repository. Specific peptide evidence found in all four databases covered a core set of a little less then 5000 genes which represent around 20% of the protein-coding genome. Around 7800 (35%) of all genes were covered by gene-specific peptides found in at least three of the four databases and ~11500 (53%) genes are matched by specific peptides from at least two databases. For the remaining 3678 genes, peptides were reported by only one of the resources, with peptides from PRIDE contributing the most additional coverage followed by PeptideAtlas, gpmDB and HuPA.

On protein level, coverage was very similar to gene level if all tryptic peptide matches, specific as well as unspecific, were considered. However, as previously mentioned, because of sequence similarities between splice isoforms, protein-specific peptide evidence was present for only ~30% of all

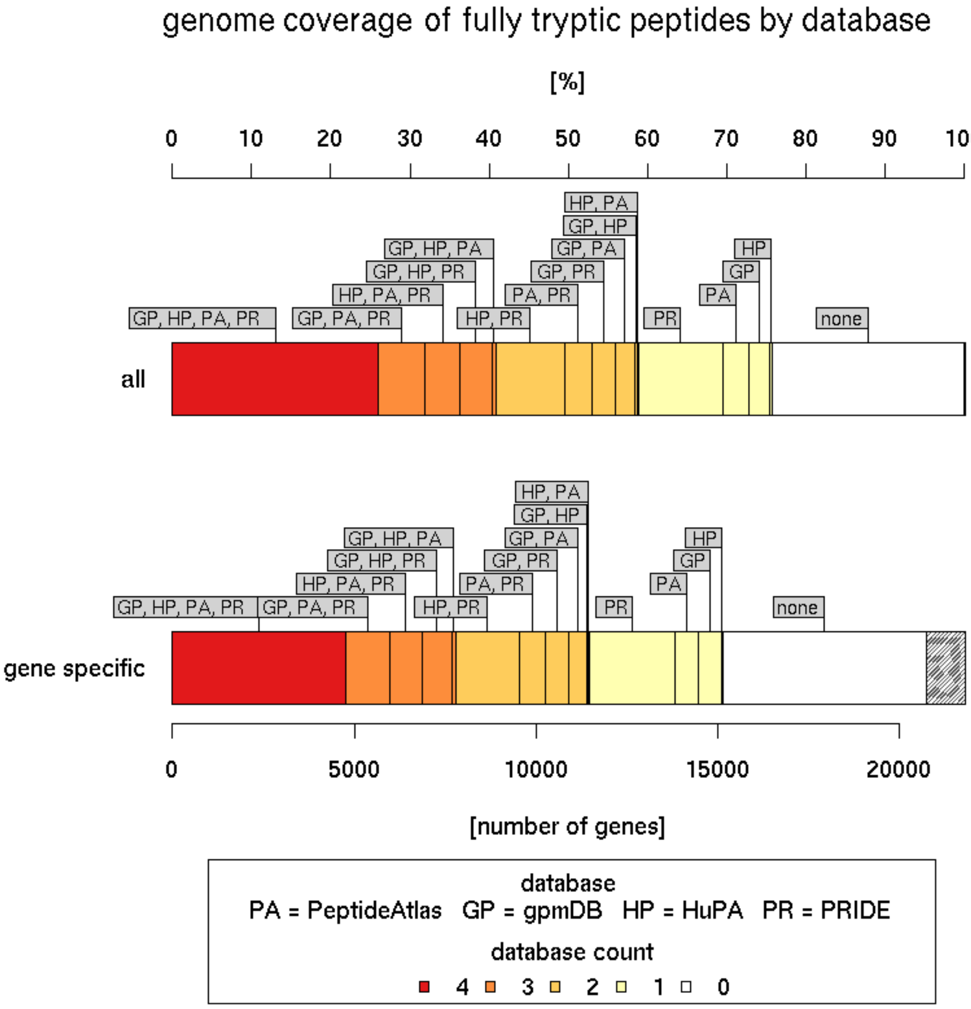


Figure 3.6: Genome coverage of fully tryptic peptides by database. Shown is the proportion of genes covered by fully tryptic peptide evidence in the four databases PeptideAtlas, gpmDB, HuPA and PRIDE. The coverage overlap between the different databases is shown by the colour coding which indicates the number of databases containing peptide evidence for the respective genes. The shaded area shows the proportion of the genome not identifiable gene-specific tryptic peptides.

protein sequences. If only sequences that emit protein-specific tryptic peptides are taken into account (~80% of all protein sequences) specific peptide evidence has been seen for ~40% of all proetin sequences uniquely identifiable by tryptic peptides.

3.3.3.3 Confirmatory evidence for coding SNPs

Peptide identifications obtained from PRIDE and HuPA are based on a range of different databases such as Ensembl, IPI, NCBInr and UniProtKB. These databases differ in the level of sequence redundancy as well as the sequence information representing a protein entity. Protein entries in the different databases can be derived from translations of mRNA or expressed sequence tag (EST) sequences carrying different alleles of a cSNP. Therefore, it is theoretically possible that peptides originating from the same position in a protein but that have been identified using different databases, carry different alleles of cSNPs found at that peptide position. A second potential source of SNP information on peptide level is provided by peptides that carry SNPs which, depending on the genotype, render the sequence identical to peptides found elsewhere in the genome. The observed sequence could then either originate from the SNP-less genome location or from the SNP-containing region, assuming the presence of the respective SNP allele.

In both peptide populations, identified and unidentified tryptic peptides, the vast majority of peptides carried no cSNP at all or less then one SNP per ten residues; 93% in case of observed peptides and 92% in case of unobserved peptides. Most observed and unobserved peptides that carried a cSNP had a SNP rate of one to two cSNPs per ten residues. The SNP rate of observed theoretical peptides was only slightly, but statistically significantly (Kolmogorov-Smirnov-Test: $D = 0.0709, p = 2.2 \cdot 10^{-16}$), lower than the SNP rate of unobserved peptides. A possible explanation could be that peptides carrying more SNPs are more likely to remain unidentified as the likelihood is higher that their sequence differs from the canonical database sequence used for spectrum matching.

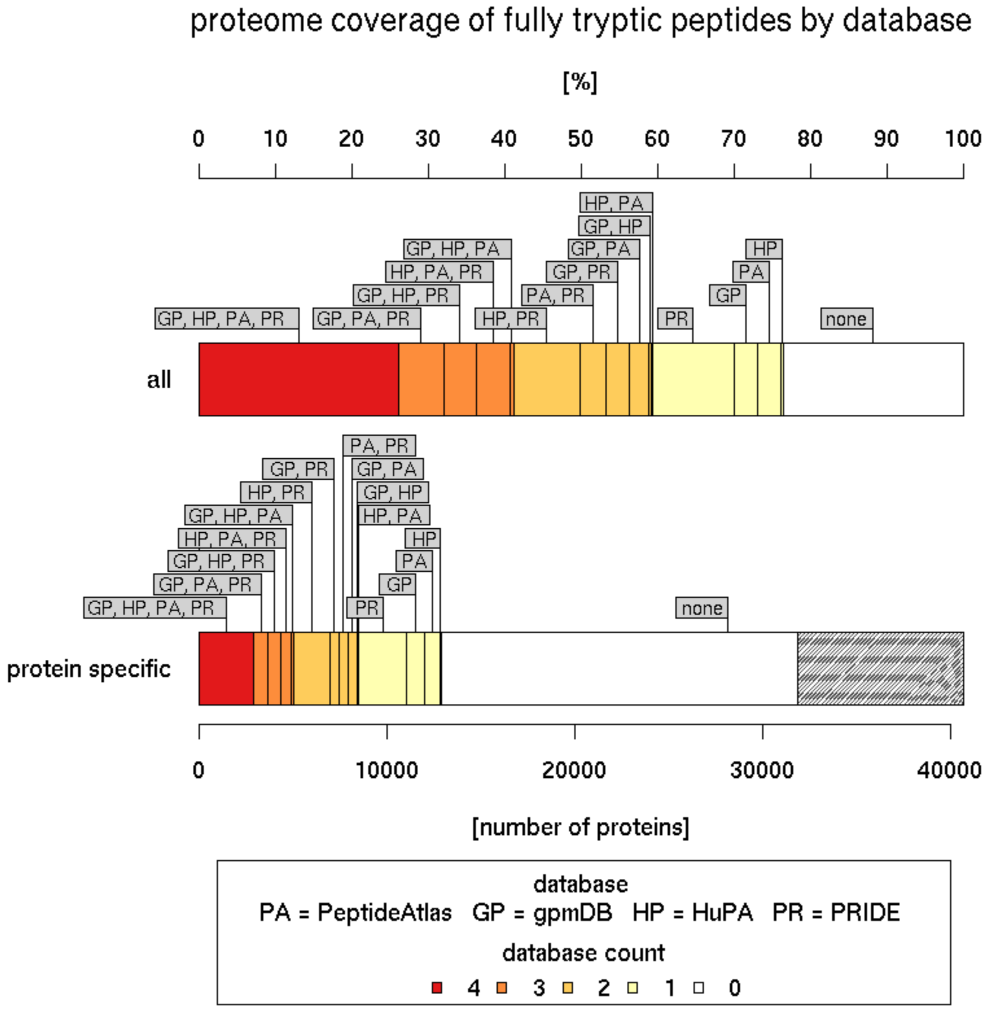


Figure 3.7: Proteome coverage of fully tryptic peptides by database. Shown is the proportion of protein sequences covered by fully tryptic peptide evidence in the four databases PeptideAtlas, gpmDB, HuPA and PRIDE. The coverage overlap between the different databases is shown by the colour coding which indicates the number of databases containing peptide evidence for the respective proteins. The shaded area shows the proportion of the proteome not identifiable by protein-specific tryptic peptides.

Together, identified tryptic peptides cover 11674 SNP locations, which is 20% of all SNP locations spanned by theoretical tryptic peptides. To test the assumption that confirmatory evidence for cSNPs on protein level could be obtained from peptide identifications made from different databases (that for the same protein contained sequence entries carrying alternative alleles of a cSNP), peptide residues at positions affected by cSNPs were compared to possible consequences of known SNP alleles.

Comparison of the SNP alleles of matched peptides with SNP data in Ensembl showed that almost all identified peptides featured amino acid residues at SNP positions that are found in the canonical Ensembl protein sequence. Only one SNP location was matched by a peptide set that covered two of the four protein level consequences at the SNP position. The sequence stretch W[A/V]AVVVPSGEEQR covered by the three peptides was found in protein products of 17 Ensembl genes and the high degeneracy of the sequence suggested that the SNP association was spurious. Closer inspection, however, revealed that all genes were located in the major histocompatibility complex (MHC) on chromosome six. The MHC locus is the most gene dense region in the human genome, with a high level of allelic diversity. Genes in the MHC locus encode proteins involved in immune response. The three genes containing the SNP with the dbSNP [275] identifier rs2308488 encode MHC class I α chains, which are subunits of protein complexes involved in antigen presentation. The remaining matched genes also encoded MHC class I proteins. The 17 genes were associated with the canonical Ensembl genome sequence annotation as well as with two different haplotype annotations of the MHC available from the Ensembl database as part of the human haplotype project [379]. Six genes were part of the regular Ensembl genome annotation, six were contributed by the annotation of the COX cell line haplotype, and five by the QBL cell line. The three SNP-containing genes were different versions of the *HLA-B* gene each associated with a different haplotype annotation, i.e. canonical, COX and QBL. Table 3.8 lists the respective Ensembl entries and figure 3.8 shows the SNP-containing protein sequence matched by peptide evidence.

Table 3.8: *HLA-B* genes containing a coding SNP covered by experimental peptide evidence. The haplotype column specifies the cell line of the haplotype annotation the respective Ensembl entries are associated with.

Ensembl gene ID	Ensembl protein ID	gene symbol	haplotype
ENSG00000204525	ENSP00000365402	*HLA-B-7*	-
ENSG00000206450	ENSP00000352656	*HLA-B-18*	QBL
ENSG00000206341	ENSP00000372816	*HLA-B-8*	COX

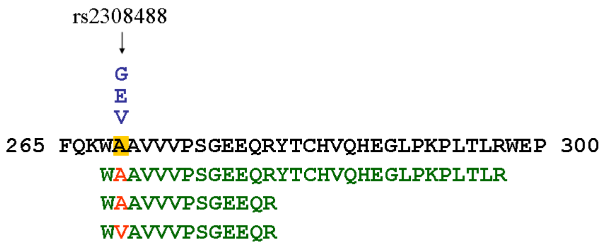

rs2308488
G
E
V
265 FQKWAAVVVPSGEEQRYTCHVQHEGLPKPLTLRWEP 300
WAAVVVPSGEEQRYTCHVQHEGLPKPLTLR
WAAVVVPSGEEQR
WVAVVVPSGEEQR

Figure 3.8: Peptide evidence for coding SNP in *HLA-B* genes. The shown stretch of sequence is found in protein sequences translated from three Ensembl gene entries of the *HLA-B* gene listed in table 3.8. The numbers to the left and to the right show the coordinates of the sequence stretch in the protein sequence. The location of the cSNP rs2308488 is shown in yellow. The alternative alleles are shown above the SNP location in blue. The three peptide sequences covering the SNP containing sequence location are shown below in green with the SNP location highlighted in red.

3.3.4 Redundancy of peptide identifications

3.3.4.1 Peptide redundancy on gene and genome level

Whether or not a peptide is observed in an MS experiment depends on various factors including its location in the protein sequence, its physicochemical properties and its abundance. In particular the latter affects the composition of MS datasets which usually exhibit a high level of peptide redundancy due to the oversampling of highly abundant proteins.

Redundancy of peptide identification on the global level was analysed based on peptide information from the PRIDE and HuPA databases. PeptideAtlas and gpmDB were not included in the analysis as they did not provide information on observation frequencies together with the proteotypic peptides. Only gene-specific fully tryptic peptides were taken into account in the analysis. The combined set of gene-specific peptides from the two databases contained 90688 different peptides covering 13213 protein-coding genes.

The box plots in figure 3.9 summarise the distributions of peptide observation frequencies on gene level. The first box plot shows the frequency distribution of peptides that are the most frequently observed peptide of a given gene. For 50% of all covered genes the most frequent peptide has been observed no more than three times. The highest non-outlier observation frequency is 20 times. A group of 1752 outliers have observation frequencies between 21 and 8226 times. The most frequently observed peptide is featured in two of the three splice isoforms encoded by the Ribosyldihydronicotinamide dehydrogenase (*NQO2*) gene. The second box plot summarises the distribution of the mean peptide observation frequencies per gene. Three quarters of all genes had an average peptide observation frequency between one and four. The gene with the highest average peptide frequency of 1508 was the Haemoglobin subunit δ (*HBD*). The last box plot shows the frequency with which the majority of a gene's peptides were observed which was between one and three.

The peptide frequency distributions indicated that a large proportion of peptide identifications reported in the two databases originated from proteins encoded by a relatively small set of genes. The extent of peptide redundancy is illustrated in figure 3.10: 50% of all peptide records in HuPA and PRIDE originated from only 85 genes, which accounted for 0.65% of all genes covered by tryptic peptide evidence from the two repositories and 0.39% of all protein-coding genes in Ensembl. A set of 1273 genes (10% of covered and 5.8% of all protein-coding genes) constituted the source of 90% of all peptide records. The remaining 11940 genes contributed a mere 10% of peptides records. On peptide sequence level, 50% of all distinct peptide sequences originated from 10% of all covered genes and 90% from half of all genes ($\sim$6600) with peptides reported in PRIDE and HuPA.

Table 3.9 lists the 85 genes 50% of all peptide records obtained from HuPA

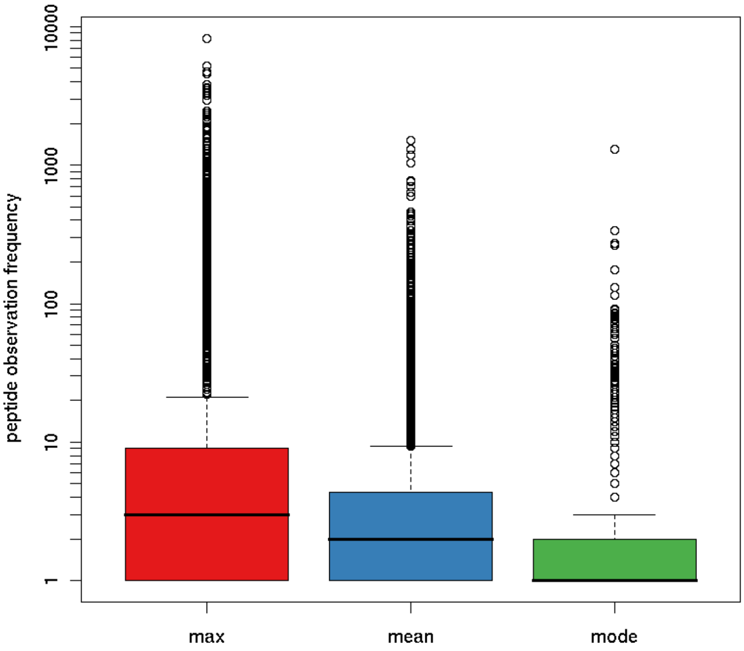


Figure 3.9: Peptide identification frequency on gene level. The box plots summarise the distributions of the maximum peptide observation frequency per gene as well as the distributions of different location measures of peptide observation frequencies.

and PRIDE originate from. A large proportion of genes are involved in the organisation of the cytoskeleton and in intracellular signaling cascades. The latter encode mostly kinases. The majority of the remaining genes are involved in housekeeping functions like amino acid, lipid and nucleotide metabolism, transcription, translation, protein degradation and intracellular transport. A number of genes function in the cell cycle. A few are associated with blood related functions such as haemostasis and complement activation. The largest contribution of peptide records by a single gene was made by titin (*TNN*), which encodes a structural protein highly abundant in skeletel muscle cells that is also the largest known protein.

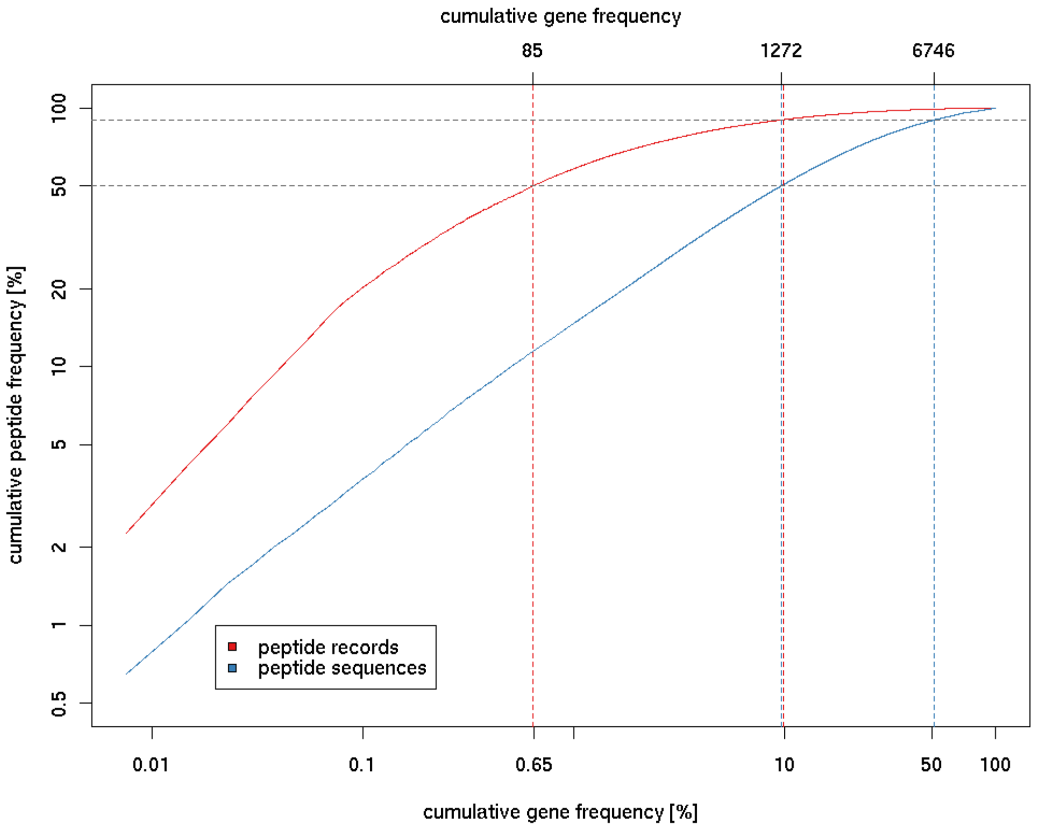


Figure 3.10: Redundancy of peptide identifications on genome level. The plot shows the cumulative peptide record and nr peptide sequence frequency vs the frequency of genes identified by the respective peptides. Both axes are log scaled.

Table 3.9: Genes featuring the most frequently reported peptides in the PRIDE and HuPA databases. Listed are the 85 genes that contributed 50% of all peptide identifications reported in the two databases. cum. pep. freq. = cumulative peptide record frequency

Ensembl gene ID	gene symbol	gene name	biological process	cum. pep. freq. [%]
ENSG00000155657	*TTN*	Titin isoform novex-3	Cytoskeleton organization	2.3
ENSG00000084674	*APOB*	Apolipoprotein B-100 precursor	Lipid metabolic process	4.3
ENSG00000121031	*PRKDC*	DNA-dependent protein kinase catalytic subunit	DNA Repair	6.0
ENSG00000197102	*DYNC1H1*	Cytoplasmic dynein 1 heavy chain 1	Mitotic cell cycle	7.7
ENSG00000125730	*C3*	Complement C3 precursor	Complement activation	9.3
ENSG00000054654	*SYNE2*	Nesprin-2	Cytoskeleton organization	10.9
ENSG00000112159	*MDN1*	Midasin	Regulation of protein complex assembly	12.4
ENSG00000178209	*PLEC1*	Plectin-1	Cytoskeleton organization	13.9
ENSG00000151914	*DST*	Bullous pemphigoid antigen 1, isoforms 6/9/10	Cytoskeleton organization	15.4
ENSG00000089154	*GCN1L1*	Translational activator GCN1	Translation	16.9
ENSG00000197694	*SPTAN1*	Spectrin alpha chain, brain	Cytoskeleton organization	17.9
ENSG00000196367	*TRRAP*	Transformation/transcription domain-associated protein	Transcription	19.0
ENSG00000149311	*ATM*	Serine-protein kinase ATM	DNA Repair	20.0

continued on next page

Table 3.9 – continued from previous page

Ensembl gene ID	gene symbol	gene name	biological process	cum. pep. freq. [%]
ENSG00000163631	*ALB*	Serum albumin precursor	Lipid metabolic process	21.0
ENSG00000175899	*A2M*	Alpha-2-macroglobulin precursor		21.8
ENSG00000198793	*FRAP1*	FKBP12-rapamycin complex-associated protein	Regulation of cell growth	22.7
ENSG00000115414	*FN1*	Fibronectin precursor	Cell adhesion	23.5
ENSG00000152818	*UTRN*	Utrophin	Cytoskeleton organization	24.2
ENSG00000196924	*FLNA*	Filamin-A	Cytoskeleton organization	25.0
ENSG00000084774	*CAD*	CAD protein	Nucleotide metabolic process	25.7
ENSG00000066279	*ASPM*	Abnormal spindle-like microcephaly-associated protein	Mitotic cell cycle	26.5
ENSG00000162434	*JAK1*	Tyrosine-protein kinase JAK1	Intracellular signaling cascade	27.2
ENSG00000085511	*MAP3K4*	Mitogen-activated protein kinase kinase kinase 4	Intracellular signaling cascade	27.8
ENSG00000183091	*NEB*	Nebulin	Cytoskeleton organization	28.5
ENSG00000115306	*SPTBN1*	Spectrin beta chain, brain 1	Cytoskeleton organization	29.1
ENSG00000120899	*PTK2B*	Protein tyrosine kinase 2 beta	Intracellular signaling cascade	29.7
ENSG00000136068	*FLNB*	Filamin-B	Cytoskeleton organization	30.2
ENSG00000091513	*TF*	Serotransferrin precursor	Iron transport	30.8
ENSG00000163359	*COL6A3*	Collagen alpha-3	Cytoskeleton organization	31.4
ENSG00000096696	*DSP*	Desmoplakin	Developmental process	31.9

continued on next page

Table 3.9 – continued from previous page

Ensembl gene ID	gene symbol	gene name	biological process	cum. pep. freq. [%]
ENSG00000100345	*MYH9*	Myosin-9	Cytoskeleton organization	32.4
ENSG00000169710	*FASN*	Fatty acid synthase	Lipid metabolic process	33.0
ENSG00000124942	*AHNAK*	Neuroblast differentiation-associated protein AHNAK	Developmental process	33.5
ENSG00000137076	*TLN1*	Talin-1	Cytoskeleton organization	33.9
ENSG00000178950	*GAK*	Cyclin G-associated kinase	Mitotic cell cycle	34.4
ENSG00000127603	*MACF1*	Microtubule-actin cross-linking factor 1, isoforms 1/2/3/5	Cytoskeleton organization	34.9
ENSG00000097007	*ABL1*	Proto-oncogene tyrosine-protein kinase ABL1	Mitotic cell cycle	35.4
ENSG00000142798	*HSPG2*	Basement membrane-specific heparan sulfate proteoglycan core protein precursor	Lipid metabolic process	35.8
ENSG00000173821	*RNF213*	RING finger protein 213		36.2
ENSG00000021826	*CPS1*	Carbamoyl-phosphate synthase	Amino acids metabolic process	36.7
ENSG00000000971	*CFH*	Complement factor H precursor	Complement activation	37.1
ENSG00000047457	*CP*	Ceruloplasmin precursor	Iron transport	37.5
ENSG00000105397	*TYK2*	Non-receptor tyrosine-protein kinase TYK2	Intracellular signaling cascade	37.9
ENSG00000078018	*MAP2*	Microtubule-associated protein 2	Cytoskeleton organization	38.2

continued on next page

Table 3.9 – continued from previous page

Ensembl gene ID	gene symbol	gene name	biological process	cum. pep. freq. [%]
ENSG00000095015	*MAP3K1*	Mitogen-activated protein kinase kinase kinase 1	Intracellular signaling cascade	38.6
ENSG00000145362	*ANK2*	Ankyrin-2	Cytoskeleton organization	39
ENSG00000171560	*FGA*	Fibrinogen alpha chain precursor	Haemostasis	39.4
ENSG00000119638	*NEK9*	Serine/threonine-protein kinase Nek9	Mitotic cell cycle	39.7
ENSG00000010671	*BTK*	Tyrosine-protein kinase BTK	Intracellular signaling cascade	40.1
ENSG00000127481	*UBR4*	E3 ubiquitin-protein ligase UBR4	Ubiquitin cycle	40.4
ENSG00000138119	*FER1L3*	Myoferlin	Cytoskeleton organization	40.8
ENSG00000175054	*ATR*	Serine/threonine-protein kinase ATR	DNA Repair	41.1
ENSG00000183735	*TBK1*	Serine/threonine-protein kinase TBK1	Intracellular signaling cascade	41.4
ENSG00000055955	*ITIH4*	Inter-alpha-trypsin inhibitor heavy chain H4 precursor	Aminoglycan metabolic process	41.7
ENSG00000106804	*C5*	Complement C5 precursor	Complement activation	42.1
ENSG00000128829	*EIF2AK4*	Eukaryotic translation initiation factor 2-alpha kinase 4	Translation	42.4
ENSG00000136628	*EPRS*	Bifunctional aminoacyl-tRNA synthetase	Translation	42.7
ENSG00000141367	*CLTC*	Clathrin heavy chain 1	Intracellular protein transport	43
ENSG00000169398	*PTK2*	Focal adhesion kinase 1	Intracellular signaling cascade	43.3
ENSG00000143322	*ABL2*	Tyrosine-protein kinase ABL2	Cytoskeleton organization	43.6
ENSG00000117983	*MUC5B*	Mucin-5B precursor	Cell adhesion	43.8

continued on next page

Table 3.9 – continued from previous page

Ensembl gene ID	gene symbol	gene name	biological process	cum. pep. freq. [%]
ENSG00000148180	*GSN*	Gelsolin precursor	Cytoskeleton organization	44.1
ENSG00000186716	*BCR*	Breakpoint cluster region protein	Mitotic cell cycle	44.4
ENSG00000111799	*COL12A1*	Collagen alpha-1	Cell adhesion	44.6
ENSG00000128815	*WDFY4*	Uncharacterized protein C10orf64.	Transport	44.9
ENSG00000127914	*AKAP9*	A-kinase anchor protein 9	Intracellular signaling cascade	45.2
ENSG00000165219	*GAPVD1*	GTPase-activating protein and VPS9 domain-containing protein 1	Intracellular signaling cascade	45.5
ENSG00000133392	*MYH11*	Myosin-11	Transport	45.7
ENSG00000147507	*LYN*	Tyrosine-protein kinase Lyn	Intracellular signaling cascade	46.0
ENSG00000166147	*FBN1*	Fibrillin-1 precursor	Haemostasis	46.2
ENSG00000151422	*FER*	Proto-oncogene tyrosine-protein kinase FER	Intracellular signaling cascade	46.5
ENSG00000166483	*WEE1*	Wee1-like protein kinase	Mitotic cell cycle	46.8
ENSG00000171564	*FGB*	Fibrinogen beta chain precursor	Haemostasis	47.0
ENSG00000140575	*IQGAP1*	Ras GTPase-activating-like protein IQGAP1	Intracellular signaling cascade	47.3
ENSG00000197893	*NRAP*	Nebulin-related-anchoring protein	Cytoskeleton organization	47.5
ENSG00000124203	*ZNF831*	Zinc finger protein 831		47.8
ENSG00000156218	*ADAMTSL3*	ADAMTS-like protein 3 precursor		48.0

continued on next page

Table 3.9 – continued from previous page

Ensembl gene ID	gene symbol	gene name	biological process	cum. pep. freq. [%]
ENSG00000165458	*INPPL1*	Phosphatidylinositol-3,4,5-trisphosphate 5-phosphatase 2	Cell adhesion	48.3
ENSG00000072518	*MARK2*	Serine/threonine-protein kinase MARK2	Intracellular signaling cascade	48.5
ENSG00000123384	*LRP1*	Prolow-density lipoprotein receptor-related protein 1 precursor	Lipid metabolic process	48.8
ENSG00000151655	*ITIH2*	Inter-alpha-trypsin inhibitor heavy chain H2 precursor	Aminoglycan metabolic process	49.0
ENSG00000110799	*VWF*	von Willebrand factor precursor	Haemostasis	49.3
ENSG00000096968	*JAK2*	Tyrosine-protein kinase JAK2	Intracellular signaling cascade	49.5
ENSG00000115464	*USP34*	Ubiquitin carboxyl-terminal hydrolase 34	Ubiquitin cycle	49.7
ENSG00000133026	*MYH10*	Myosin-10	Cytoskeleton organization	50.0

3.3.4.2 Functional spectrum of genes covered at different levels of peptide redundancy

To identify potential biases introduced by the redundancy in peptide and protein identification, the functional composition of gene sets at different levels of peptide coverage was analysed. Protein-coding genes were grouped into four sets: i) genes covered by the 50th percentile ($P50$) of peptide records, ii) genes covered by peptide records between the 50th and the 90th percentile ($P50 - 90$), iii) genes covered by peptide records between the 90th to the 100th percentile ($P90 - 100$) and iv) genes without peptide matches ($P0$). Genes were then categorised based on their association with terms of the Gene Ontology (GO) *molecular function* ontology and over- and under-representation of functional categories assessed based on the hypergeometric P value of the observed term frequency.

Figure 3.11 shows the functional classification of genes at each level of coverage. Shown is the distribution of genes across the 'generic' Slim subset of the GO *molecular function* ontology. Orange bars show the frequency of protein-coding genes in Ensembl and purple bars the frequency of genes in the respective set annotated with a given GO term. Bars to the right show the log_2 ratio of the term frequencies in the gene set and Ensembl. The ratio of significantly over-represented terms are shown in green and the ratio of significantly under-represented terms in red. The significance threshold was established for each set by controlling the false discovery rate (FDR) at 0.01%.

The functional categorisation of the 85 genes contributing half of all peptide records discussed in the previous section (3.3.4.1) already suggested an enrichment of cytoskeletal and cytoskeleton associated proteins as well as kinases in the $P50$ gene set. The Gene Ontology analysis of the $P50$ set showed that the enrichment of the respective categories - actin binding, structural molcule activity, cytoskelal protein binding, transferase activity, kinase activity, protein kinase activity, enzyme regulator activity - is indeed statistically significant. Significantly under-represented on the other hand were genes encoding nucleic acid binding proteins including tran-

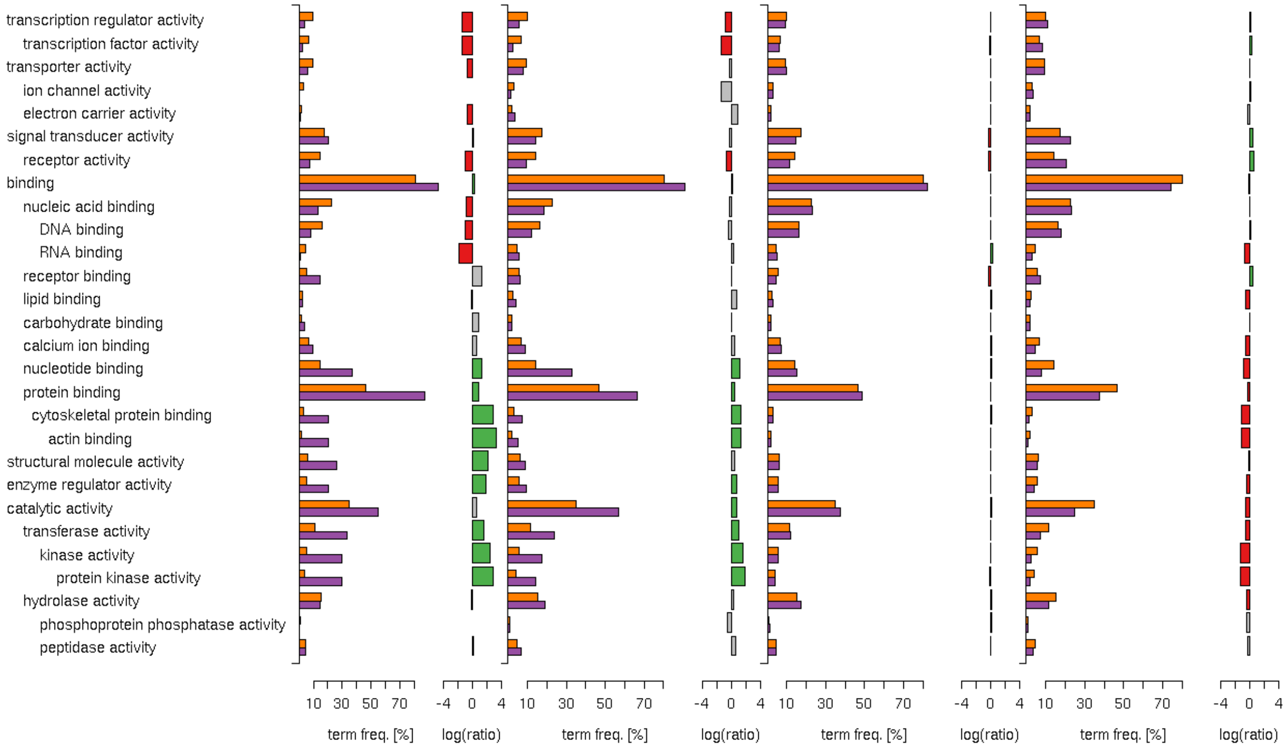


Figure 3.11: Functional categorisation of genes covered at different levels of peptide redundancy . Shown is the distribution of genes across the Slim subset of the GO *molecular function* ontology. Orange bars show the frequency of protein-coding genes in Ensembl and purple bars the frequency of genes in the set annotated with the respective term. Only terms associated with at least 10% of all protein-coding genes in Ensembl are included. Bars to the right show the $log_2(term\ frequency\ gene\ set/term\ frequency\ Ensembl)$. Ratios of significantly over-represented terms are shown in green, ratios of significantly under-represented terms in red. The significance threshold for over- and under-representation was established for each set by controlling the FDR at 0.01%.

scription factors, as well as functional categories associated with transmembrane proteins such as receptor and transporter activity.

The biases observed in the $P50$ gene set were largely found in the $P50-90$ set as well. Though structural molecules were no longer significantly enriched; however, cytoskeleton associated proteins still were. Furthermore, the under-representation of nucleic acid binding proteins *per se* was no longer significant. However, transcription factors were still significantly under-represented.

Of all four gene sets the functional composition of the $P90-100$ fraction was least biased and came closest to the overall composition of the protein-coding genome, though genes with receptor activity were still under-represented as were genes encoding receptor binding proteins.

The functional categorisation of the $P0$ gene set showed that the undetected part of the protein-coding genome is enriched in functions under-represented in the covered part. This includes genes involved in transcription regulation, transmembrane receptors and receptor binding proteins.

3.3.4.3 Transcript expression of genes covered at different levels of redundancy

The analysis of the Human Proteome Organisation (HUPO) Brain Proteome Project (BPP) dataset showed that frequently identified proteins tend to be encoded by mRNAs with higher expression levels. To verify if this observation also holds true on a global level mRNA expression levels from the GNF Atlas of Gene Expression were examined for the four gene sets identified at different levels of redundancy. The box plots in figure 3.12 summarise the distribution of median expression values across all tissues and cell lines in the GNF Expression Atlas for genes in the respective sets.

Compared to all transcripts interrogated in the GNF Atlas (red) the subset mappable to Ensembl protein-coding genes (blue) showed slightly increased expression values. The overall distribution of expression values of Ensembl genes matched by tryptic peptides from PRIDE or HuPA (green) was similar to that of Ensembl protein-coding genes in general with a slight shift towards higher expression levels observable. Although the median expression level of genes in the $P50$ set (purple) was equal to that of the distribution of covered genes (green), the frequency of genes with higher expression values was increased. This trend was continued in the $P50-90$ set (orange) which showed the clearest shift towards higher expression levels. Expression levels fell again in the $P90-100$ set (yellow) and were lowest in the uncovered set of genes (brown) which showed a distribution very similar to the overall distribution of the GNF dataset.

Similar to the observations in the analysis of the HUPO BPP dataset frequently identified proteins tended to be encoded by genes with higher expression levels. The correlation was not quite as clear as for the HUPO BPP data, probably because the comparison here was based on the median expression across all tissues, and expression levels show a high tissue-dependent variance. In case of the HUPO BPP dataset, on the other hand, the expression data was filtered for a specific tissue.

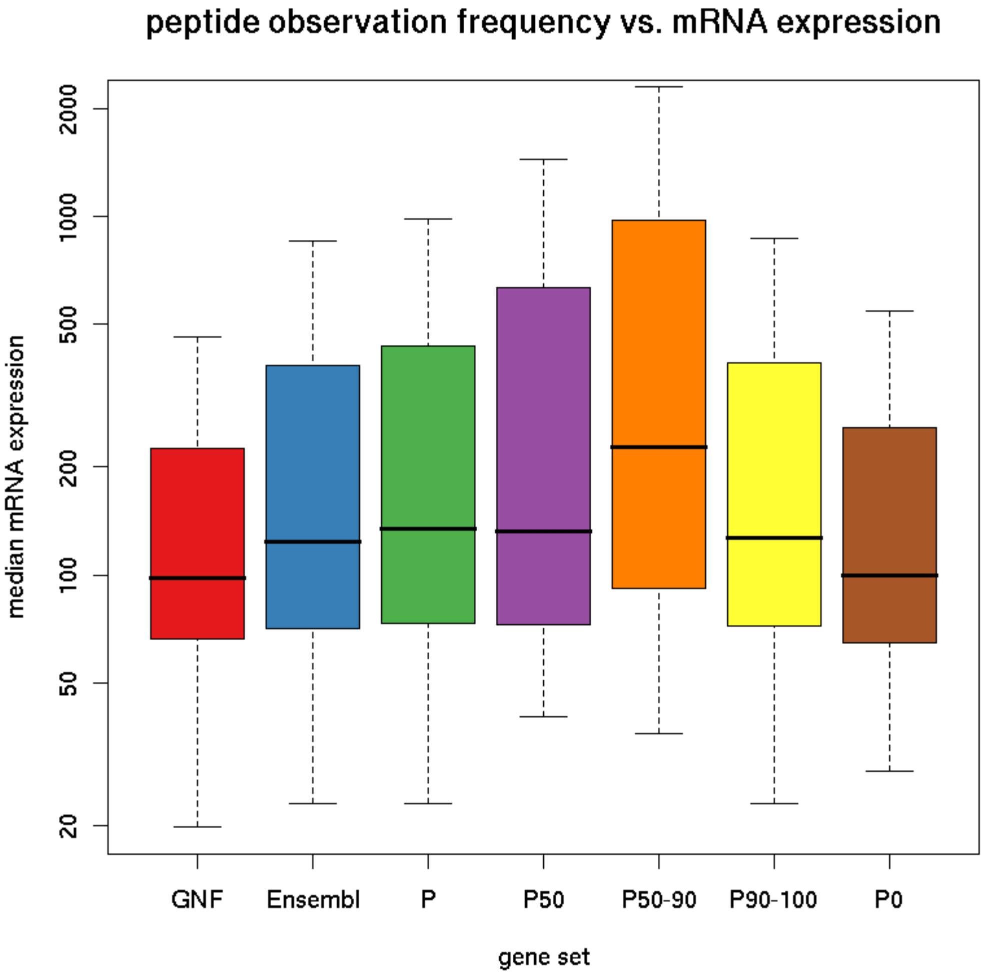


Figure 3.12: Transcript expression levels of genes covered at different levels of peptide redundancy. The box plots summarise the distribution of median expression values across all tissues and cell lines in the GNF Atlas of Gene Expression for genes in the respective set. GNF = all transcripts interrogated in the GNF atlas, Ensembl = Ensembl protein-coding genes, P = all genes matched by at least one tryptic peptide from PRIDE or HuPA, $P50/P50-90/\ P90-100$ = genes covered by the respective percentiles of peptide records (see section 3.3.4.1), $P0$ = genes without peptide matches.

3.4 Discussion

3.4.1 Concordance of expression evidence for protein-coding genes obtained from different public proteomics resources

The comparison of peptide data available from four major public proteomics repositories showed that despite similar numbers of non-redundant peptide sequences contained in the databases, the overall overlap of sequence information is relatively small. Several factors could explain the complementary nature of peptide information. Firstly, peptide identifications in the different databases were made based on data generated by different experimental approaches in various different sample types, with limited overlap in sample origins and approaches between the databases. Genes and isoforms expressed vary across different tissues and thus the potentially observable peptides. Likewise, which peptide of a protein is observed depends on the ionisation method and MS technology used. Secondly, the limited reproducibility of mass spectrometry experiments *per se*, due to the stochastic variability of peptide observability, could be responsible for the differences in peptide information contained in each database. Thirdly, different algorithms were used for spectral matching which each have their biases. Furthermore, a comparison based directly on peptide sequence identity might be too stringent a similarity measure as the same sequence region of a protein might emit different proteolytic peptides resulting from missed cleavage or truncation of peptides. The latter point was addressed by assessing the database overlap based on theoretical tryptic peptides covered by observed fully tryptic peptide sequences, thereby "normalising" the data with regard to proteolytic heterogeneity. Although the relative overlap increased, the overall agreement remained with 12% still relatively low.

These results illustrate the need for an integration of available proteomics data to reach the full potential of proteomics information in the public domain. This is of particular importance when it comes to achieving the goal that PeptideAtlas and gpmDB have set out to accomplish: the identification of proteotypic peptide reporters and selected reaction monitoring (SRM) signatures for targeted proteomics, based on empirical data [376, 53, 380, 381]. Currently no single resource contains sufficient data to achieve coverage of the empirical proteotypic peptide space anywhere close to that achieved by all databases combined, and the computational prediction of such peptides [339] has proved challenging [382].

Efficient integration of proteomics data generated around the world will be of particular importance if a systematic and coordinated effort to map the human proteome is not realised and information generated in a decentralised way has to be exploited effectively.

3.4.2 The use of mass spectrometry proteomics data for genome annotation

Cleavage products of trypsin normally observed in proteomics experiments are fully tryptic [54] and peptides with tryptic termini are the most abundant components of proteolytic digests with trypsin [55]. Furthermore, database searches often penalise non-tryptic cleavage, resulting in higher probabilities being assigned to fully tryptic peptides. This results in an under-representation of non-tryptic peptides in databases even if samples might contain a high number of unspecifically cleaved peptides emitted from highly abundant proteins. The peptide population analysed here also predominantly contained peptides with tryptic termini and only fully tryptic peptide evidence was considered for the analysis of confirmatory evidence for genome annotation.

Because of the high stringency of trypsin specificity it has been suggested previously that only fully tryptic peptides should be used for protein identification [54]. The results obtained here show that doing so does not compromise gene or protein coverage to a significant extent. This is true for both specific as well as degenerate peptide evidence.

Compared to the peptide level, the overlap between the databases increases on gene and protein level with almost a third of all genes covered by gene-specific peptide evidence from all four databases. This observation is consistent with previous results showing that proteomics experiments become more comparable on the level of protein identifications [383]. This makes sense, as the heterogeneity in proteotypic peptide observations resulting from different sample types, technologies and search algorithms is “normalised” on protein and gene level.

The fact that individual databases provided peptide evidence not found in any other database for a quarter of all identified genes yet again underlines the synergistic effect of combining data from different resources. A similar effect was observable on the level of protein-specific peptides, i.e. peptides that match exactly one protein in the proteome (as opposed to gene-specific peptides matching only protein products of one gene). While peptides found in all four databases could unambiguously identify a mere 7% of all protein sequences, the combined set of specific peptides covers more than 31% of the proteome. The coverage of unambiguous expression evidence is even more remarkable if the absence of unique tryptic peptides for around 20% of all protein sequences is taken into account. The combined coverage of the uniquely identifiable proteome then amounts to 41%.

In comparison to the UniProtKB statistics, the analysis of genome coverage provided a slightly more optimistic estimate of expression evidence for protein-coding genes on protein level, yielding an additional 15% of covered genes. It has to be taken into account, however, that the analysis presented here considered all available peptide data without applying a quality filter. The ’protein existence’ annotation provided by UniProtKB on the other hand, is the result of a stringent manual curation process.

The high coverage of the protein-coding genome by peptide evidence highlights the considerable progress MS based proteomics has made since an earlier systematic analysis of peptide data in the context of the human genome [149]. Yet, the results also underline the shortcomings of current

MS proteomics data with regard to genome and proteome annotation going beyond the pure verification of a gene being translated into a protein product. Fine-grained evidence for gene models is relatively sparse, and genes with full exon and splice event coverage tend to have relatively simple gene models. Only a handful of alternatively spliced genes had confirmatory evidence on the isoform level. Furthermore, the analysed peptide data was, apart from one example, devoid of confirmatory evidence for coding SNPs. On nucleotide level SNP information has been obtained by mining sequence databases [384], demonstrating that the information is contained in the primary databases. In theory this information should be carried through to protein sequence databases, however, in practice most databases of CDS translations used for spectrum identification seem to be homogenised regarding sequence variation as a result of redundancy removal which suppresses small variations. IPI, for example, clusters sequences at 95% sequence similarity and contains one representative master entry for each sequence cluster. Other databases, like NCBInr, group sequence entries only if they are 100% identical and could thus potentially harbour SNP variants. However, such databases are not commonly used in proteomics because the increased database size results in an increased run-time of identification programmes and can negatively influence the statistical significance a search engine assigns to an identification. EST databases have been used to overcome the limitations of protein sequence databases regarding isoforms [385, 386, 387], however, protein sequence searching remains the predominant method of spectrum identification in proteomics. An alternative strategy is the enrichmment of protein sequence databases with SNP information. This solution has been implemented in the MSIPI database, a modified version of IPI that contains protein sequences elongated with additional peptides for the identification not only of coding SNPs but also N-terminal signal and transit sequences [388].

3.4.3 Biases in the functional spectrum of mass spectrometry proteomics data

The small proportion of the protein-coding genome covered by 90% of all identified peptides emphasises the extent of the problem caused by the redundancy of peptide identifications. Frequently identified proteins tend to be highly and ubiquitously expressed and often already well characterised proteins. Mostly these proteins are involved in the organisation of the cytoskeleton and in basic metabolism.

Depletion of highly expressed proteins is one solution to overcome the masking of low-abundance proteins and is routinely used in proteomic analyses of plasma, which shows a particularly large dynamic range [389, 390]. Given the available information on identification redundancy and the relatively small group of proteins emitting the majority of repeatedly identified peptides, a similar approach should be applicable to other sample types. This assumes the availability of the required affinity reagents. Efforts to systematically generate antibodies for human proteins are underway [391, 231] and antibodies have been produced already for a considerable proportion of human proteins [392]. Still, there are caveats to a depletion approach. One problem is the potential co-precipitation of untargeted proteins that bind to the depleted proteins [352]. Furthermore, once the first sequence of highly abundant proteins is removed the next series of most abundant proteins will cause similar problems for the detection of low abundance proteins [390].

While abundance is one reason for the biases in the functional spectrum of proteomics data, another important determinant are the physicochemical properties of proteins associated with certain functions. This is the case in particular for transmembrane receptors and transcription factors. Standard proteomics approaches select against both classes because of their hydrophobic or basic character. In these cases strategies tailored to the respective protein classes are required and have already been established [393, 394].

Another strategy to overcome biases of current MS based proteomics is the use of targeted approaches for peptide identification. Currently most proteomics experiments are based on data-dependent acquisition, which selects peptides for fragmentation based on precursor ion intensity. In targeted approaches based on SRM the targeted peptides are detected specifically by isolating precursors with predefined m/z ratios and monitoring the occurence of specific precursor-product-ion combinations uniquely identifying the targeted peptide. Targeted MS based approaches based on SRM will be the subject of the next chapter.

Eventually it will probably be a combination of improved sample preparation techniques and more targeted approaches for peptide identification that will help to overcome current biases in MS based proteomics. For proteomics experiments in general this is essential to achieve a higher coverage of proteins important for the understanding of the specific system, phenotype or pathology under study. For the systematic mapping of the human proteome overcoming current shortcomings is required to achieve the necessary resolution and coverage for a fine-grained annotation of genes and gene models based on experimental evidence at protein level. The latter in turn provides the basis for robust targeted identification of proteins, as verified gene models will be a requirement for the selection of suitable peptide reporters.

3.4.4 Conclusions and perspectives

The results presented in this chapter reflect on a global level the trends observed on study level in the analysis of the HUPO BPP dataset. Current mass spectrometry proteomics data is highly redundant as a result of the predominant identification of abundant proteins and selection against proteins with certain physicochemical properties. As a result proteomics datasets are biased in their functional composition and offer limited peptide coverage on the gene model and SNP level. This emphasises the need for a systematic effort to map human proteins using targeted MS approaches that can specifically identify those peptides crucial for confirmation of gene models and coding SNPs on the protein level. Additionally, this approach can also deliver expression evidence for genes that are currently without experimental evidence on the protein level.

This also touches on the topic of whether a HPP, and proteomics in general, should be gene- or protein-centric [374]. The discussion surrounding this topic seems to confuse the question of whether the proteome should be studied on protein, transcript or genome level, with the question of whether proteomics data should be analysed in the context of the genome sequence. MS based proteomics by definition is protein-centric since measurements are made on protein level. Nonetheless, experimental design and a meaningful interpretation of the results also have to take into account the genomic context.

Genomic context matters if the proteome is to be studied systematically on a large scale. The target selection for splice event verification, for example, requires knowledge of transcript models. The same applies to the identification of isoform specific peptide reporters.

Genomic context should also play an important part in the interpretation of any MS proteomics experiments in general and in particular in eukaryotic systems that exhibit a high level of alternative splicing. The majority of studies report minimal explanatory sets of proteins for the observed peptide evidence. The results presented here show that currently the majority of alternative isoforms cannot be resolved based on the identified peptides. Therefore, if different isoforms of an alternatively spliced gene share the same peptide evidence, it is often the case that the isoforms reported as identified have been selected based on criteria not necessarily biologically meaningfull such as their annotation status or length. However, given the insufficient resolution of splice isoforms, a more appropriate approach for the time being might be an interpretation of the results on the gene level.

Gene level interpretation of results neglects, of course, the domain modularity of genes and the functional differences as well as the tissue dependent expression of splice isoforms. However, gene-centric interpretation of the results does not falsely assume a granularity on isoform level that is not backed by experimental evidence.

Finally, using the annotated genome sequence as a reference for the systematic mapping of the proteome makes sense, as it provides a natural coordinate system for protein level data. Genome annotation pipelines use many lines of evidence to delineate the protein-coding genome, resulting in a more robust estimation of the protein sequence space. Agreeing on one reference sequence space should also help to overcome, at least to some extent the "Bable" of proteomics resulting from the heterogeneity of sequence databases used for protein identification. Using the annotated genome sequence as a reference would eliminate the confusion caused by different levels of redundancy and coverage in the different databases. Although, the problem of propagating data across different versions of genome assembly and annotation would still remain.

So the answer to the question of whether proteomics should be gene or protein-centric should be that it necessarily has to be both, not least if information from the genome project is to be utilised effectively in the future. Given the poor resolving power of current MS proteomics approaches, however, a gene-centric interpretation of results is probably more approriate at this time.

Chapter 4

Conquering uncharted territory: estimation of the scope and selectivity of a targeted proteomics strategy based on combinatorial proteolysis

4.1 Introduction

Shotgun or bottom-up proteomics is one of the most powerful and widely used mass spectrometry (MS) based proteomics approaches to analyse complex mixtures of proteins. A protein sample prepared from a biological source is digested with a protease, commonly trypsin. The resulting peptide mixture is then separated by liquid chromatography, the peptides ionised and analysed by tandem mass spectrometry. First, a mass spectrum of peptide or precursor ions is recorded and the most abundant precursor ions are selected individually for fragmentation. Then a mass spectrum of the resulting fragment or product ions is recorded which is used to obtain peptide sequence information.

Despite considerable technological and methodological advances in recent years, the complexity and dynamic range of proteomes still pose major challenges to the global analysis of protein expression by shotgun proteomics. Firstly, selection of precursor ions for fragmentation based on peak intensity, referred to as data-dependent acquisition, results in a bias towards peptides of highly abundant proteins and poor sensitivity for petides of proteins with low expression levels. The impact of this bias on the composition of proteomics datasets has been discussed in detail in the preceeding chapters. Secondly, the automated selection of precursor ions leads to inconsistent reproducibility of peptide identifications by shotgun proteomics. Thirdly, for any given protein only a small number of proteolytic peptides are repeatedly and consistently identified [339]. Often, however, these so-called *proteotypic* peptides are degenerate peptides, i.e. peptides that match more then one protein sequence - usually different isoforms of the same gene - and are thus not suitable for unambiguous protein identification [382]. Again this has been elucidated in the previous chapters.

These shortcomings led to the emergence of more targeted proteomics strategies. Instead of globally profiling the expression of all proteins in a sample, only a defined set of proteins of interest is monitored [41]. This is achieved through the targeted detection of protein-specific peptide reporters by selected reaction monitoring (SRM), a highly specific and sensitive mass spectrometry method. A mass-spectrometer, typically a triple-quadropole instrument, run in SRM mode will monitor the occurance of precursor ions with specific m/z values defined in an inclusion list. The detection of a monitored mass will trigger the fragmentation of the respective precursor. The resulting product ions are in turn monitored for the occurance of a specific product ion m/z value. The combination of precursor and product ion m/z values is referred to as a transition, and one or a combination of several transitions can uniquely identify a certain peptide sequence.

It has been demonstrated that the detection of sequence specific proteotypic peptides by SRM improves the detection of those protein classes that are under-sampled in shotgun proteomics experiments [395] and enables more accurate quantification of homologous proteins [396]. SRM has been employed successfully to detect proteins with high sensitivity in plasma [397, 92], monitor sites of post-translational modification [398, 399, 93, 400, 401] and quantify biomarkers [402, 403].

Although SRM is still not widely used, the promise of the technology is being recognised by the proteomics community, and tools to support the design and management of SRM experiments are becoming available [380, 381, 404, 405].

Starting from a list of target proteins, the design of an SRM experiment involves the establishment of a set of suitable peptide reporters that uniquely identify the respective protein sequences and are detectable by MS. Observable fragment ions that discriminate the targeted peptide from the background of isobaric peptides present in the sample must then be identified for each reporter peptide.

Designing SRM transitions is a heuristic process that is laborious and time consuming. Empirical data on the proteotypic character and fragmentation of peptides in proteomics repositories can be used to help streamline the process [380, 381, 404]. However, as shown in the previous chapter the proteome coverage of such data is limited in particular with regard to sequence specific peptides. Algorithms to computationally predict proteotypic peptides based on empirical data have been published [339, 406, 407]. Yet, for a large number of proteins, no proteotypic peptides could be predicted [339]. Furthermore, a comparison of predicted and experimentally observed peptides suggests a limited performance of such predictions [382]. This emphasises the need for unbiased methods to identify candidate peptide reporters and transitions to facilitate the discovery and experimental validation of SRM transitions.

Apart from transition design, another factor impacting upon the success of targeted proteomics approaches is the availability of specific peptide reporters for the targeted set of proteins. Traditionally, peptides detected in MS based proteomics experiments are generated by proteolysis with trypsin. Trypsin is the protease of choice in proteomics experiments firstly because of its high specificity [55, 54], and secondly because cleavage products carry basic residues (lysine or arginine) at the carboxy (C)-terminal end which is advantageous for peptide sequencing based on collision induced dissociation (CID). However, as pointed out in chapter 3, trypsin fails to generate sequence-specific peptides for around 20% of all proteins encoded in the human genome, limiting the scope of targeted proteomics approaches.

It has been demonstrated that the use of proteases other than trypsin has beneficial effects on protein identification. Combining peptide identifications from complementary peptide populations generated by multiple proteases increases the number of peptides detectable by tandem MS, resulting in a significant increase in sequence coverage [59, 408, 409, 58, 410, 411, 57]. Diversification of the peptide population by combinatorial proteolysis could thus be a suitable strategy to increase the number of proteotypic peptide reporters detectable by mass spectrometry.

In this chapter the performance of a tryptic digest in generating a peptide population beneficial for the unambiguous identification of proteins in a targeted proteomics experiment is compared to that of a combinatorial strategy. Different combinations of proteases are scored based on a cost-benefit analysis that takes into account the complexity of the resulting peptide mixtures and the proteome coverage by sequence-specific peptide reporters. Finally, the detectability of peptide reporters in the generated mixture by SRM is estimated by an exhaustive search of theoretical spectra for minimal sets of candidate transitions to unambiguously identify targeted peptides against the background of isobaric peptides.

4.2 Materials & methods

4.2.1 Protein sequences

Translations of protein-coding transcripts of Ensembl (release 45) were obtained in FASTA format from the Ensembl file transfer protocol (FTP) server (ftp://ftp.ensembl.org/pub/release-45/homo_sapiens_45_36g/data/fasta/pep/Homo_sapiens.NCBI36.45.pep.all.fa.gz). A non-redundent sequence database was constructed by grouping entries with identical amino acid sequence resulting from translations of coding sequences encoding the same amino acid sequence.

4.2.2 *In silico* proteolytic digest

DBToolkit (version 3.5.4) [378] was used to *in silico* digest the non-redundent FASTA sequence database with the proteases Arg-C, Lys-C, V8 (V8D/V8DE), pepsin A, and trypsin, not allowing for missed cleavages. Table 4.1 lists the protease cleavage site definitions used. Note that the specificity of the V8 protease is dependend on the presence/absence of phosphate. In digestion buffers that contain phosphate the protease cleaves after glutamic and aspartic acid (V8DE). In the absence of phosphate the protease is specific to glutamic acid (V8E) residues.

Table 4.1: Protease cleavage site definitions. The proteases all cleave C-terminal of the specified residue(s) (cleaved) if not followed by a restricting residue (restricts).

protease	cleaved	restricts
Trypsin	K,R	P
Pepsin A	F,L	-
Arg-C	R	P
Lys-C	K	P
V8E	E,Z	P
V8DE	B,D,E,Z	P

4.2.3 Algorithm to determine combinations of SRM transitions uniquely identifying a targeted peptide

The algorithm used to identify combinations of product ion peaks in the spectrum of targeted peptides that uniquely identify a peptide against a background of isobaric peptides in a peptide population generated by a given protease is summarised by the pseudo code in algorithm 1.

Algorithm 1 Pseudo-code explaining the algorithm used to search product ion spectra for peak combinations uniquely identifying a target peptide against a background of isobaric peptides

```
$a$ := mass tolerance

for all target peptides $t$ do

  create empty set of target product ion combinations $C_E \leftarrow \emptyset$

  calculate target precursor ion mass-to-charge ratio $MZ(precursor_t)$

  select set of background peptides $B_t$ with each background peptide $b \in B_t$
  having $|MZ(precursor_t) - MZ(precursor_b)| \leq 2 \cdot a$

  for all combinations $C_i$ of product ions $product_1, ..., product_n \in$ target spec-
  trum $s_t$ do

    select set of background peptides $B_e$
    excluded by product ion combination $C_i$
    with each excluded background peptide spectrum $s_e \pm a \not\supset C_i \pm a$

    if $B_e = B_t$ then
      add combination $C_i$ to target product ion combinations $C_E$
    end if

  end for

end for
```

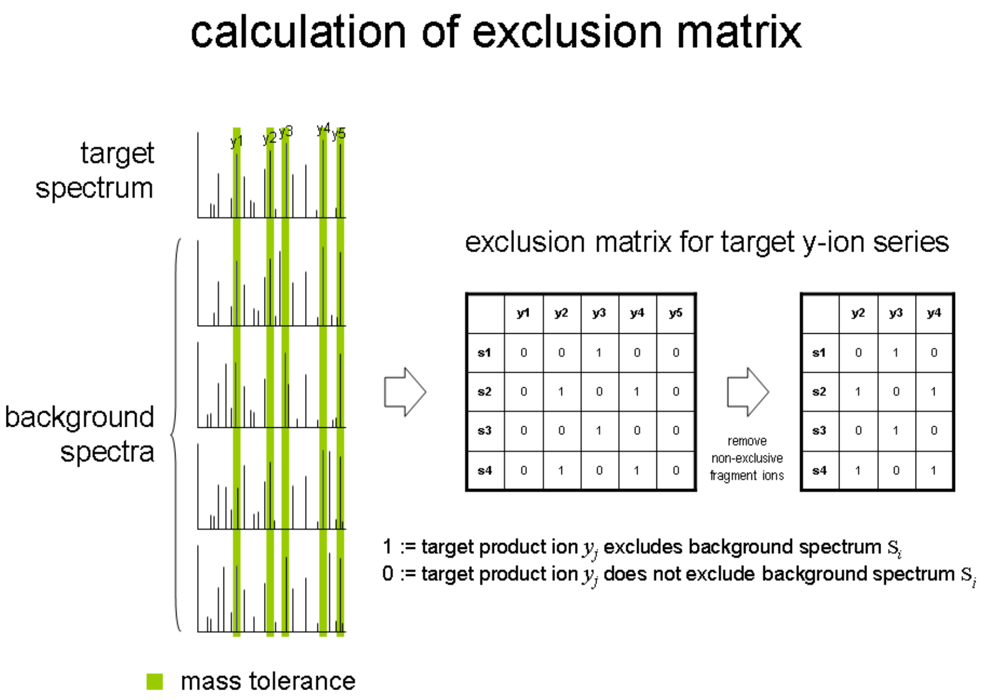
For a given target peptide the set of background peptides is calculated by selecting all background peptides with mass-to-charge ratios overlapping with that of the target peptide, given a specified mass tolerance and a set of allowed precursor ion charge states.

The product ion spectrum of the targeted peptide is then searched for product ion combinations that are not found in any of the background peptide spectra, given the specified mass tolerance and the allowed product ion charge states. Which product ion types are considered is user defined. By default $y-$ions of the target peptide spectrum are searched against a background of $y-$ and $b-$ions.

Starting with the smallest combination size the algorithm iterates over all possible combinations of target product ions of a given size until either the first unique combination is found or all unique combinations of that size are found, which depends on the user defined settings. If no unique combination is found the combination size is incremented and the search repeated. This procedure is repeated until either one or more unique combinations or all combinations have been tested.

The search for unique target product ion combinations is performed on an $m \times n$ 'exclusion' matrix that is calculated for each target peptide and the respective background peptides. Figure 4.1 illustrates the calculation of the matrix. The number of rows m of the matrix equals the number of background peptides, and the number of columns n of the matrix equals the number of target product ions considered in the search. A cell a_{ij} of the matrix has a value of 1 if the spectrum s_i of background peptide i is excluded by the target product ion j. This means the spectrum of background peptide i does not contain a product ion with an m/z value that overlaps with that of target product ion j given the mass tolerance and allowed product ion charge states.

For each combination of target product ions that is tested for exclusion of background spectra, an exclusion vector for the respective matrix columns is calculated containing the maximum row values of the sub-matrix (see figure 4.2). If the sum of vector values equals the number of background peptides, the combination excludes all background peptides and the combination is added to the set of excluding product ion combinations.


calculation of exclusion matrix
target spectrum
y1
y2
y3
y4
y5
background spectra
exclusion matrix for target y-ion series
remove non-exclusive fragment ions
🟩 mass tolerance

	y1	y2	y3	y4	y5
s1	0	0	1	0	0
s2	0	1	0	1	0
s3	0	0	1	0	0
s4	0	1	0	1	0

	y2	y3	y4
s1	0	1	0
s2	1	0	1
s3	0	1	0
s4	1	0	1

1 := target product ion y_j excludes background spectrum s_i

0 := target product ion y_j does not exclude background spectrum s_i

Figure 4.1: Identification of signature transitions: calculation of the exclusion matrix. Hypothetical example illustrating the calculation of an exclusion matrix for a target product ion spectrum and a set of background product ion spectra of isobaric peptides.

4.2.4 Calculation of peptide, precursor and product ion masses

The theoretical molecular mass M_P of a neutral peptide P with n_a amino acid residues was calculated using equation 4.1 where A_i is the monoisotopic residue mass of the neutral amino acid at position i, N is the molecular mass of the neutral amino (N)-terminal group and C the mass of the neutral carboxy (C)-terminal group.

$$M_P = M(P) = N + \sum_{i=0}^{n_a} A_i + C \tag{4.1}$$

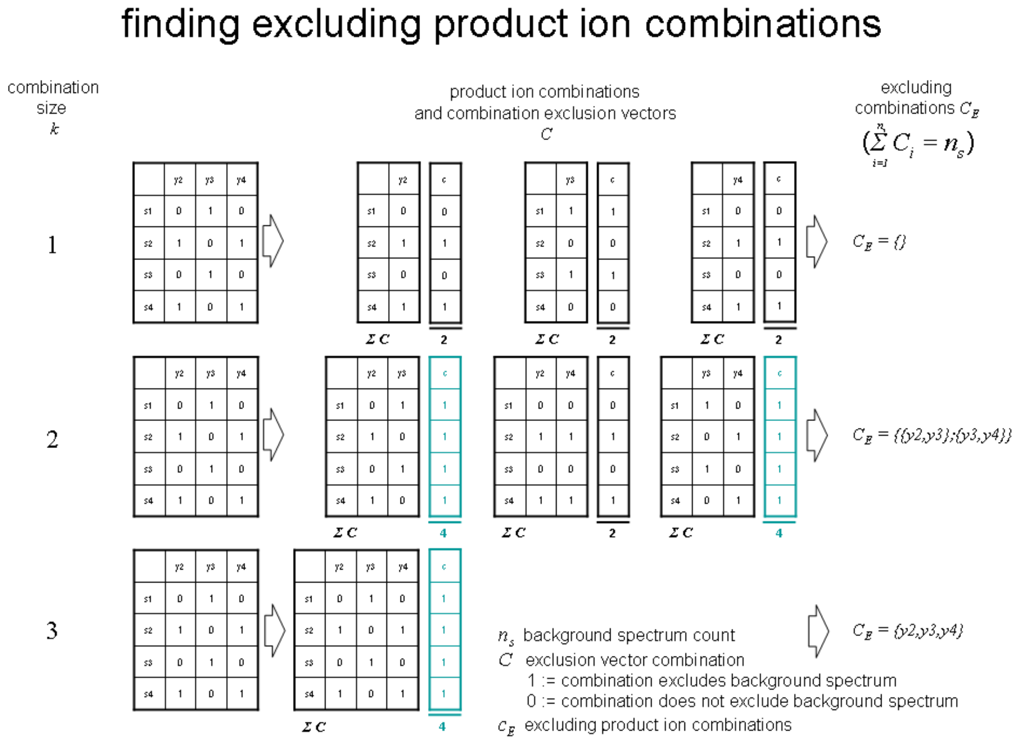

finding excluding product ion combinations
combination size k
product ion combinations and combination exclusion vectors C
excluding combinations C_E
$(\sum_{i=1}^{n_s} C_i = n_s)$
1
2
3
y2
y3
y4
s1
s2
s3
s4
c
ΣC
$C_E = \{\}$
$C_E = \{\{y2,y3\};\{y3,y4\}\}$
$C_E = \{y2,y3,y4\}$
n_s background spectrum count
C exclusion vector combination
1 := combination excludes background spectrum
0 := combination does not exclude background spectrum
C_E excluding product ion combinations

Figure 4.2: Identification of signature transitions: finding product ion combinations exclusive to the targeted peptide. For each product ion combination an exclusion vector of the respective matrix columns is calculated containing the maximum value of each row of the sub-matrix. The sum of vector values equals the number of background ion spectra for target product ion combinations that exclude all background spectra.

Molecular masses M_F of neutral peptide fragmentation products F were calculated using the formulae in table 4.2 where N is the molecular mass of the neutral N-terminal group and C the mass of the neutral C-terminal group and M_P is the molecular mass of the neutral amino acid residues and $M(x)$ the molecular mass of molecule group x.

The theoretical mass-to-charge ratio R_P of an ionised peptide P with charge z was calculated according to equation 4.2 where M_H is the monoisotopic mass of hydrogen.

Table 4.2: Formulae for the calculation of neutral product ion masses.

fragmentation product type	neutral molecular mass M_F
a	$N + M_P - M(CHO)$
a*	$a - M(NH_3)$
a°	$a - M(H_2O)$
b	$N + M_P - M(H)$
b*	$b - M(NH_3)$
b°	$b - M(H_2O)$
c	$N + M_P + M(NH_2)$
x	$C + M_P + M(CO) - M(H)$
y	$C + M_P + M(H)$
y*	$y - M(NH_3)$
y°	$y - M(H_2O)$
z	$C + M_P - M(NH_2)$

$$R_P = \frac{M_P + z \cdot M_H}{z} \tag{4.2}$$

4.2.5 Molecular masses of elements, amino acid residues and modified amino acid residues

Tables 4.3, 4.4 and 4.5 list the molecular masses of elements, amino acid residues and modified amino acid residues that were used in all calculations.

Table 4.3: Monoisotopic element masses.

element	symbol	monoisotopic element mass [Da]
Hydrogen	H	1.0078
Carbon	C	12.0000
Nitrogen	N	14.0031
Oxygen	O	15.9949

Table 4.4: Monoisotopic amino acid masses.

amino acid	one-letter code	monoisotopic residue mass [Da]
Alanine	*A*	71.0371
Arginine	*R*	156.1011
Asparagine	*N*	114.0429
Aspartic acid	*D*	115.0269
Cystein	*C*	103.0092
Glutamine	*Q*	128.0586
Glutamic acid	*E*	129.0426
Glycine	*G*	57.0215
Histidine	*H*	137.0589
Isoleucine	*I*	113.0841
Leucine	*L*	113.0841
Lysine	*K*	128.0950
Methionine	*M*	131.0405
Phenylalanine	*F*	147.0684
Proline	*P*	97.0528
Serine	*S*	87.0320
Threonine	*T*	101.0477
Tryptophan	*W*	186.0793
Tyrosine	*Y*	163.0633
Valine	*V*	99.0684

Table 4.5: Monoisotopic masses of modified amino acid residues.

modification	monoisotopic residue mass [Da]
$MetO$	147.0354
$MetO_2$	163.0303
$CysO_3$	135.0990
$TrpO_2$	218.0691

4.3 Results

4.3.1 Coverage of the proteome by non-degenerate proteolytic signature peptides

4.3.1.1 Degeneracy of tryptic peptides

The most informative peptides with regard to protein identification in an MS based proteomics experiment are those that can originate from exactly one protein sequence (given the proteome model). These peptides will be refered to as *signature peptides* in the following, thereby distinguishing them from proteotypic peptides which are peptides frequently observed across MS based proteomics experiments for a given protein but are not necessarily required to be unique to exactly one protein isoform. Trypsin is the most commonly used protease in MS based proteomics. To assess the performance of trypsin regarding the generation of signature peptides for unambiguous protein identification, the proteome-wide degeneracy of tryptic peptides was analysed based on an *in silico* digest of human protein sequences. Here, the proteome is defined as the set of non-redundant protein sequences in Ensembl.

Proteolysis of the 38644 non-redundant protein sequences in Ensembl resulted in 622688 distinct peptide sequences with masses in the optimal performance range of MS between 600 - 4000 Dalton (Da). Figure 4.3 A gives an overview of the degeneracy of the peptide population generated by trypsin. The bar plot shows that 56.9% of tryptic peptides were unique across the Ensembl protein set, while the remaining peptides could originate from two or more proteins.

4.3.1.2 Sources of peptide degeneracy

In principle there are two sources of peptide degeneracy: i) intra-gene sequence similarity, i.e. the peptide originates from a stretch of amino acid sequence encoded by an exon that is part of more than one isoform of

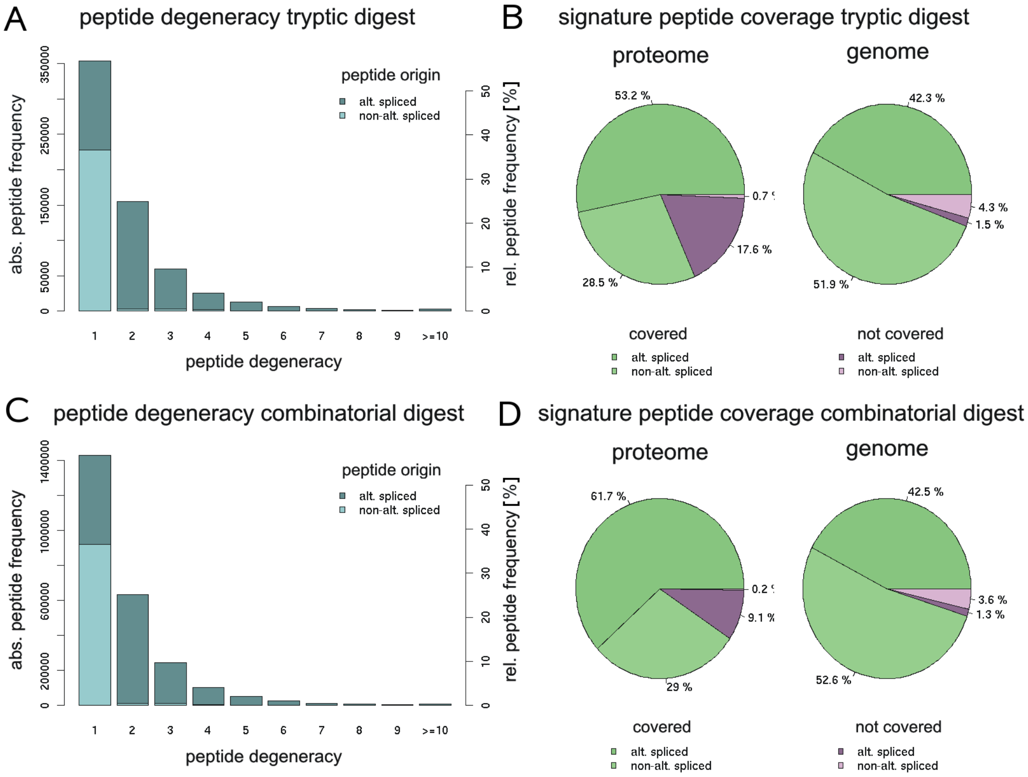


Figure 4.3: Comparison of peptide populations generated by a tryptic and a combinatorial proteolytic digests in terms of peptide degeneracy and proteome/genome coverage by signature peptides. A) Degeneracy of peptides generated by trypsin. The degeneracy of a peptide is given by the number of protein sequences it maps to. The Bar height shows the frequency of peptide sequences at each level of degeneracy. The shading indicates the proprotion of peptides originating from alternatively and non-alternatively spliced genes. B) Coverage of the proteome and genome by tryptic signature peptides (degeneracy = 1). Green areas show the proportion of protein sequences and genes covered. Purple areas show the proportion not covered. The different shadings distinguish alternatively and non-alternatively spliced protein products and genes respectively C) Degeneracy of peptides generated by a parallel digest with the five proteases trypsin, pepsin A, Lys-C, Arg-C and V8 (V8D/V8DE). D) Coverage of the proteome and genome by peptides generated by the protease combination. alt. = alternatively.

an alternatively spliced gene, or ii) inter-gene sequence similarity, i.e. the peptide originates from a stretch of amino acid sequence that is part of the coding region of more than one gene as a result of sequence homology between conserved domains or gene families.

Around 50% of protein-coding Ensembl genes (10043) encode alternative protein products. The shading of the bars in figure 4.3 A shows the proportion of peptides originating from alternatively spliced and non-alternatively spliced genes at each level of degeneracy. The majority of degenerate peptides were emitted by proteins encoded by alternatively spliced genes, indicating that intra-gene sequence similaritry between protein isoforms is the largest contributer to peptide degeneracy. That this is indeed the case is shown in figure 4.4: more than 90% of degenerate peptides were the result of intra-gene similarity between protein isoforms, while only a relatively small proportion of peptides mapped to protein products encoded by different genes. Some peptides were degenerate across both protein isoforms as well as genes.

4.3.1.3 Proteome and genome coverage of tryptic signature peptides

Figure 4.3 B shows the fractions of the proteome and genome covered by tryptic signature peptides respectively. As the green areas of the pie chart show, sequences of 81.7% of all Ensembl protein entries emitted at least one signature peptide when digested with trypsin. Of those protein entries which did not have a unique tryptic peptide (purple and pink areas), 96% were the product of alternatively spliced genes (purple areas). On genome level a peptide is considered a signature peptide if it can unambiguously detect the expression of a particular gene, i.e. the peptide matches only protein products encoded by exactly one gene. The peptide population generated by trypsin contained 587754 signature peptides uniquely identifying 94.2% of protein-coding genes in Ensembl.

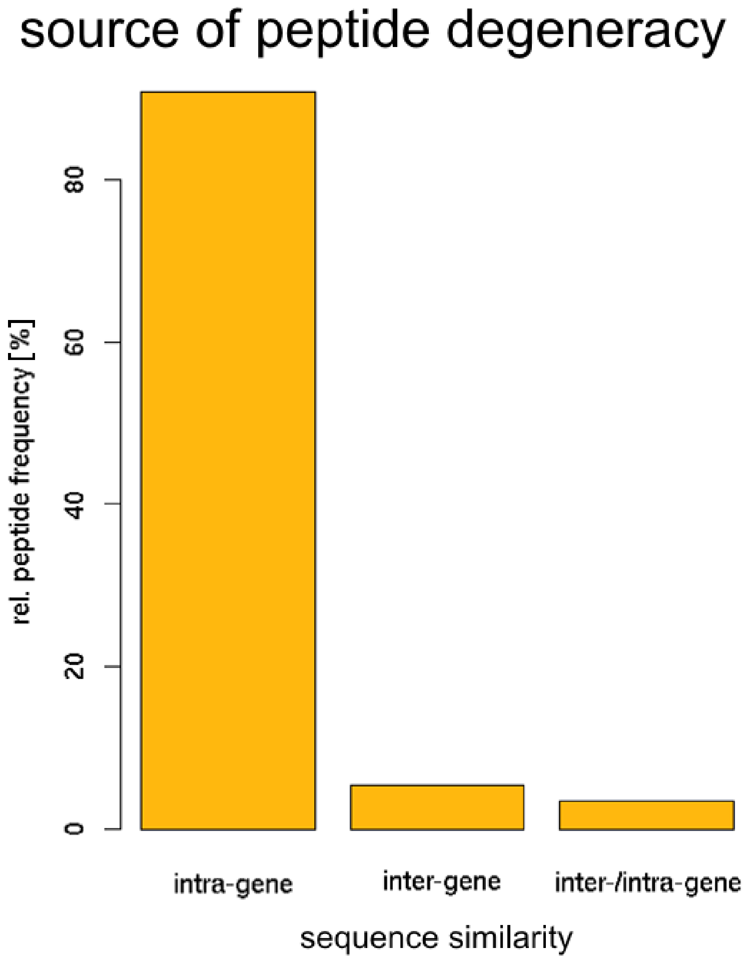


Figure 4.4: Source of peptide degeneracy. The bar plot shows the proportions of degenerate tryptic peptides resulting from inter-gene sequence similarity between splice isoforms and intra-gene sequence similarity, for example between homologous genes.

4.3.1.4 Proteome and genome coverage of signature peptides generated by a combination of proteases

Figures 4.3 B shows that the use of trypsin left 18.3% of the proteome and 5.8% of the genome inaccessible to unambiguous identification by shotgun proteomics approaches due to the absence of signature peptides. To test the hypothesis that the coverage of the proteome by signature peptides could be increased through the diversification of the analysed peptide mixture by using a combination of different proteases, the degeneracy of a peptide population generated by an *in silico* digest with five proteases was analysed. In addition to trypsin, four other proteases routinely employed in mass spectrometry experiments were used, namely pepsin A, Arg-C, Lys-C, and V8. The specificity of V8 can be modulated by adding phosphate to the digestion buffer. In presence of phosphate V8 cleaves C-terminal of both glutamic and aspartic acid (V8DE). In the absence of phosphate cleavage occurs only after glutamic acid (V8E).

The principle of the proposed combinatorial approach is illustrated again in figure 4.5. The protein sample is split into equal aliquots which are each subjected to proteolysis with a different protease in parallel. After completion of the digest the samples are combined and the resulting pool of complementary peptide populations is analysed by MS.

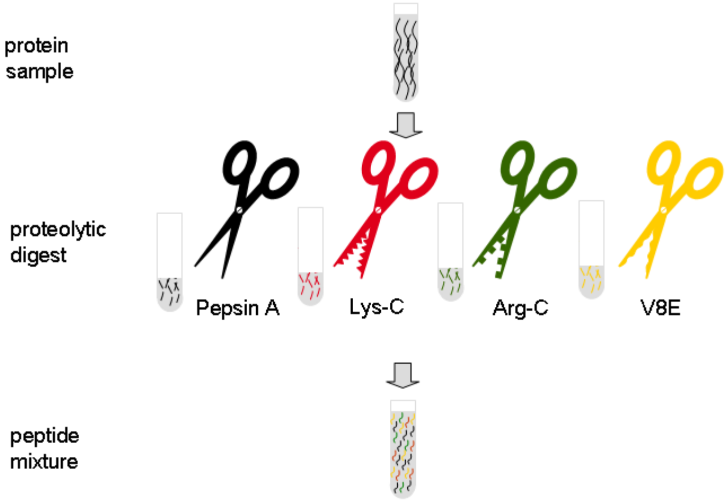


Figure 4.5: Diversification of peptide populations by parallel combinatorial proteolysis. The protein sample is split into equal aliquots each digested with a different protease to create complementary peptide populations which are subsequently pooled for mass spectrometric analysis.

The peptide population generated by the five proteases trypsin, pepsin A, Arg-C, Lys-C, V8 (V8DE and V8D) contained 2519972 distinct peptides of which 1430278 (56.8%) were unique signature peptides on protein level (see figure 4.3 C); a proportion very similar to that generated by trypsin alone. However, the fraction of the proteome identified by these peptides was 90.7% which equals an 11% increase in proteome coverage compared to trypsin alone (see figure 4.3 D). Coverage of splice isoforms increased by 16% , from 53.2% to 61.7%, meaning that an additional 6200 isoforms were resolved by signature peptides (dark green area). Coverage of the non-alternatively spliced proteome increased slightly from 28.5% to 29.0%.

On genome level the increase in coverage was not quite as dramatic as on proteome level. Compared to trypsin signature peptides were present for an additional 196 genes corresponding to a 1% increase. The biggest increased was observed for non-alternatively spliced genes.

4.3.1.5 Cost-benefit analysis of different protease combinations

Comparing the total number of peptides generated by trypsin and the combinatorial approach shows that the increase in proteome coverage by signature peptides comes at the cost of significantly increasing the complexity of the resuting peptide mixture. To identify the protease combination with the best trade-off between the increase in coverage and the increase in complexity a cost-benefit analysis of different protease combinations was carried out. The results are summarised in figure 4.6.

The top plot shows the proteome coverage by signature peptides for each of the 63 possible combinations of the six protease activities Lys-C, Arg-C, pepsin A, trypsin, V8D and V8DE. The relative frequency of signature peptides in the peptide mixture is plotted below, followed by the increase in complexity compared to a tryptic digest. Here, complexity is defined as the number of distinct theoretical peptide sequences in the mass range 600 - 4000 Da generated by an *in silico* digest with the respective protease combination. Each combination was scored by calculating the benefit-cost ratio of coverage increase Δc (given by the difference between the coverage of a the respective combination c_{comb} and that of trypsin c_{tryp}) and complexity increase i_x (measured as x-fold complexity of the tryptic peptide mixture). The bottom-most plot shows the benefit-cost score which corresponds to the $\Delta c / i_x$ ratio scaled to values between 0 and 1. The plots are ordered along the x-axis in ascending order of the combination indices shown in the table on the right. Combination indices are shown as the bottom x-axis labels. Combinations are grouped by size, with the groups separated by dotted lines. The combination size in each group is indicated by the numbers at the top of the plot. Dashed horizontal lines mark the value for the tryptic peptide population in the respective plot.

The cost-benefit analysis showed that the increase in proteome coverage by 10.9% compared to trypsin, achieved when all six proteases are combined, came at the cost of a five-fold increase in sample complexity. It also revealed that the highest proteome and genome coverage was not achieved when all six protease activities were combined, but by the combination of the five protease activities Lys-C, Arg-C, pepsin A, V8DE and V8E. Signature peptides were generated for 39260 proteins, ten proteins more than with all six proteases, at a lower complexity increase of only four-fold.

Figure 4.6 also shows that the increase in coverage started to level off relatively quickly at a combination size of two. Combinining trypsin and pepsin A already resulted in an 8.4% increase in proteome coverage at a complexity increase of 2.1-fold. Adding a third protease, either V8 cleaving after glutamic acid (V8D) or V8 cleaving after glutamic and aspartic acid (V8DE), resulted in a further improvement to a 10.2% increase, which was already very close to the maximally achievable coverage increase of 10.9 %; however, the complexity increased by only 2.8- and 3.1-fold, respectively.

Based on the benefit-cost score calculated for each protease combination the peptide population generated by the three proteases Lys-C, Arg-C and V8E achieved the best trade-off between signature peptide coverage and complexity. Table 4.6 gives details of the highest scoring combinations for each combination size. It shows that the benefit-cost ratio droped for combinations of more than three proteases, while the gain in coverage was

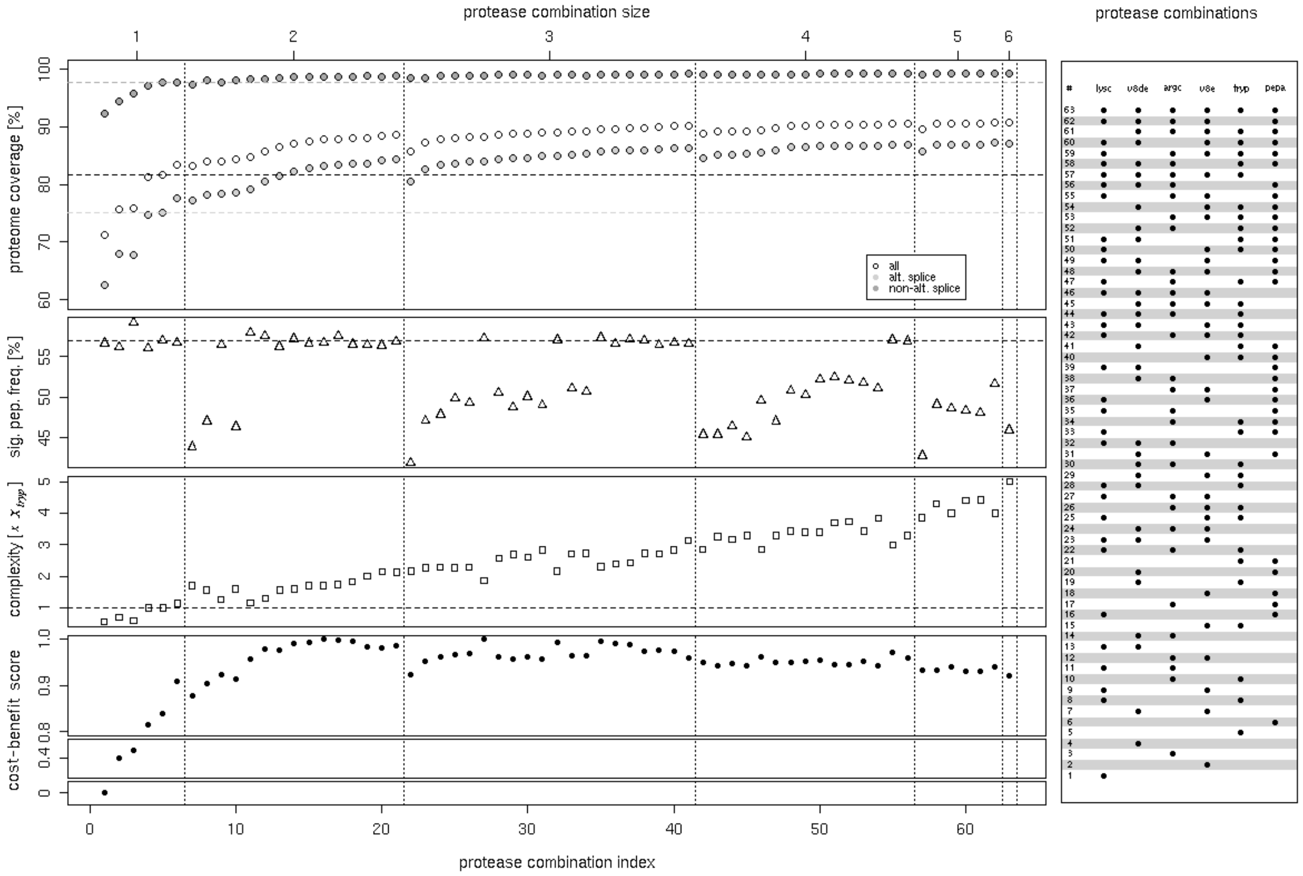


Figure 4.6: Cost-benefit analysis of protease combinations. Plotted are the proteome coverage, signature peptide frequency, complexity increase and benefit-cost score for all 63 combinations of the five proteases Lys-C (lysc), Arg-C (argc), pepsin A (pepa), trypsin (tryp) and V8 (v8d,v8de). See section 4.3.1.5 for a detailed description.

only marginal. This is explained by the fact that the specificity of the two proteases V8DE and trypsin overlaps with that of other proteases already in the combination. While trypsin combines the specificity of Arg-C and Lys-C, V8DE shares a cleavage site with V8D and recognises an additional amino acid residue (see table 4.1). These proteases generate shorter peptides which are subsets of peptides already in the mixture. As they can only be of the same or lower specificity than the longer peptides they are subsets of, these peptides do not contribute to the signature peptide coverage and only increase the complexity of the mixture.

4.3.1.6 Sequence coverage by signature peptides for different protease combinations

Combining proteases increased the number of proteins emitting signature peptides. It also increased the absolute number of signature peptides. However, this does not affect the probability of observing a signature peptide as the relative proportion of signature peptides stays more or less the same across combinations. Yet, a higher sequence coverage by signature peptides might have a positive impact on the signature peptide detectability as it increases the chance of generating a signature peptide from a sequence region that is likely to produce peptides observable by mass spectrometry.

The histograms in figure 4.7 show the sequence coverage of tryptic signature peptides in comparison to those generated by the highest scoring protease combinations in table 4.6.

Sequence coverage by signature peptides increased with the number of proteases used. Using all six proteases in combination resulted in a 32-fold increase of sequences with a coverage >95%. The protease combination with the highest cost-benefit score {Lys-C, Arg-C, V8E} still achieved a 15% increase of such protein sequences.

Plotting coverage seperately for proteins translated from alternatively spliced and non-alternatively spliced transcripts shows that the biggest proportion of proteins with more than 95% coverage was largely con-

Table 4.6: Protease combinations with the highest cost-benefit score. For each combination size the protease combination with the highest cost-benefit score on proteome level is shown. Complexity x is defined as the number of unique peptides in the mass range 600 - 4000 Da generated by a protease combination. Complexity increase i_x is measured as fold complexity x_{tryp} of the peptide mixture generated by an *in silico* tryptic digest, coverage increase Δc is given by the difference between the coverage of a the respective combination c_{comb} and that of trypsin c_{tryp}. The benefit-cost ratio is given by the coverage increase Δc normalised by the increase in complexity i_x.

protease combination	signature peptide frequency	complexity	proteome coverage	complexity increase	coverage increase	benefit cost ratio
		x	c	i_x	Δc	$\Delta c/i_x$
	[%]	[$\times 10^6$ peptides]	[%]	[$\times x_{tryp}$]	[%]	[%]
Pepsin A	56.69	0.7	83.44	1.1	2.13	1.87
Lys-C Pepsin A	56.66	1.1	87.74	1.7	7.38	4.33
Lys-C Arg-C V8E	57.26	1.2	88.31	1.9	8.08	4.34
Lys-C Arg-C V8E Pepsin A	57.04	1.9	90.46	3.0	10.71	3.57
Lys-C Arg-C V8E V8DE Pepsin A	51.61	2.5	90.67	4.0	10.97	2.74
Lys-C Arg-C V8E V8DE Pepsin A Trypsin	45.95	3.1	90.65	5.0	10.94	2.19

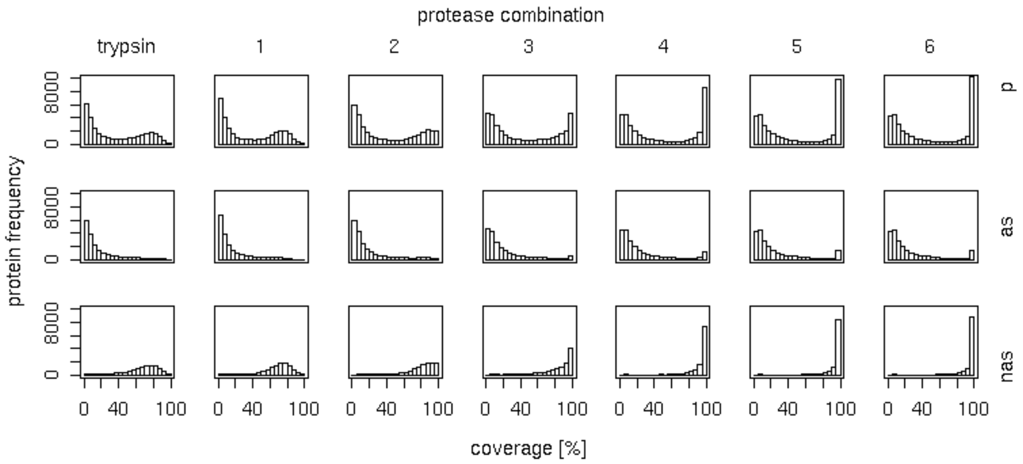


Figure 4.7: Sequence coverage by signature peptides for different protease combinations. The histograms show relative sequence coverage by signature peptides on the x-axis vs protein frequency on the y-axis for i) the whole proteome (p), ii) translations of alternatively spliced transcripts (as), iii) and translations of non-alternatively spliced transcripts (nas). Coverage is plotted for trypsin and the combinations 1 = {pepsin A}, 2 = {Lys-C, pepsin A}, 3 = {Arg-C, Lys-C, V8E}, 4 = {Lys-C, Arg-C, V8E, pepsin A}, 5 = {Lys-C, Arg-C, V8E, pepsin A, V8DE}, 6 = {Lys-C, Arg-C, V8E, pepsin A, V8DE, trypsin}.

tributed by genes with only one translation. However, there was also a clear increase of sequence coverage for products of alternatively spliced genes.

4.3.2 Detectability of signature peptides by selected reaction monitoring

Combinatorial proteolysis provides a means to increase the sensitivity of targeted proteomics experiments through the generation of signature peptides for unambiguous identification of proteins that do not emit such peptides when digested with trypsin.

The second step in a targeted proteomics workflow is the selective identification of such peptides by SRM. To investigate the feasability of selectively detecting signature peptides generated by the highest scoring protease combination {Arg-C, Lys-C, V8E} by SRM, theoretical product ion spectra were searched for peak combinations in the signature peptide spectrum unique to the targeted peptide, against the spectra of isobaric background peptides. See section 4.2.3 for details of the search algorithm. Since the identified product ion combinations could be used as a basis for the development of specific SRM transitions they are referred to as *candidate signature transition sets* in the following.

Figure 4.8 shows an example of a candidate signature transition set for the peptide ETSMVHELNR. The transition set consists of three y-ions whose combination is unique to the spectrum of the targeted peptide at the chosen mass tolerance of $\pm 1 m/z$ for both precursor and product ions. The set of y_3-, y_4- and y_6-ion uniquely identifies the target spectrum against the background y- and b-ion spectra of isobaric peptides. Peak locations of the approximately 3000 background spectra are shown as green and pink dots, with each horizontal line representing one background spectrum. The blow-ups of the mass tolerance window (grey shaded area) around the signature product ions shown below illustrate that the majority of background spectra do not contain any of the three products. The $\sim$250 spectra that do have product ions falling within the respective mass tolerance windows are excluded by the combination of ions.

4.3.2.1 Estimation of signature peptide detectability in an unmodified peptide population

The results of an exhaustive search of all signature peptide $y-$ion spectra against a background of b- and y-ions assuming a mass tolerance of $\pm 1 m/z$, a precursor ion charge of $z = +2$ and a product ion charge of $z = +1$ are summarised in figure 4.9.

The mass distribution of unmodified signature (dashed line) and total peptides (continuous line) generated by the protease combination {Arg-C, Lys-C, V8E} is shown by the green curves in the top left corner of figure 4.9 A. Similarly to a tryptic peptide population the peptide frequency

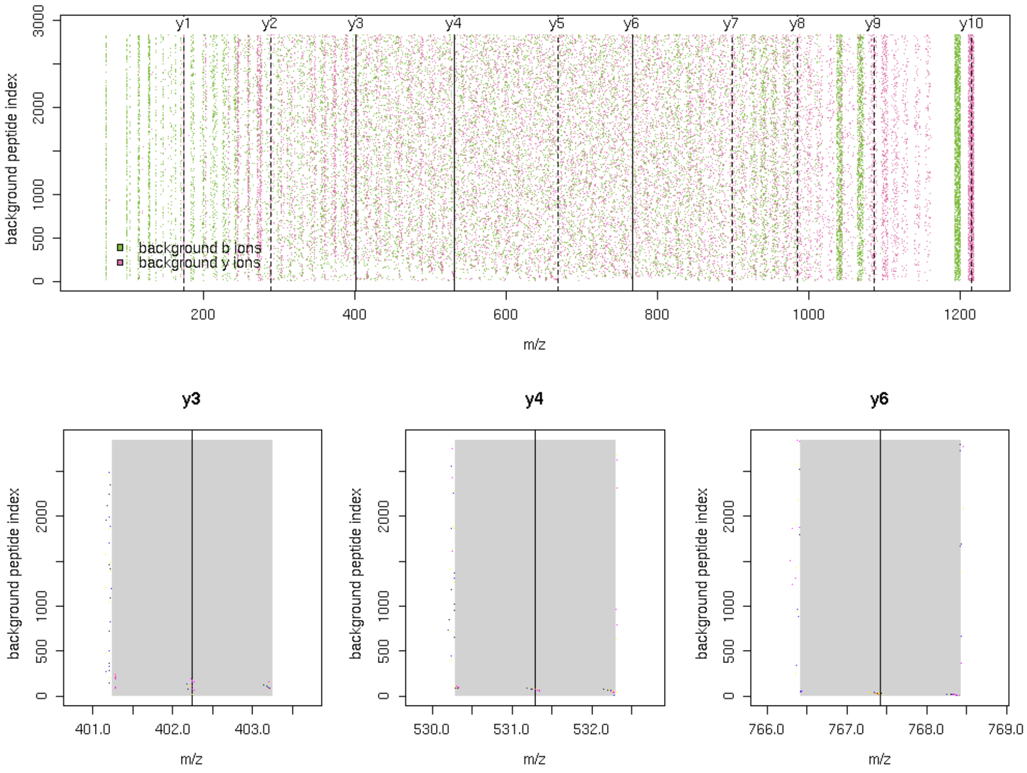


Figure 4.8: Candidate signature transition example. The top plot shows the theoretical y-ion sepctrum for peptide ETSMVHELNR (disregarding peak intensities). The three product ions making up the signature transition set are shown as continuous lines. The combination of y_3-, y_4- and y_6-ion is unique to the spectrum of the peptide against the spectra of isobaric peptides. Green and pink dots show m/z values of the product ions of the ~3000 background peptides whose precursor m/z values overlap with the target peptide given a mass tolerance of $\pm 1 m/z$. Each horizontal line represents one background spectrum. The plots at the bottom show blow-ups of the mass tolerance window around each signature product ion (grey area).

peaks at around 1000 Da and decreases steadily for higher peptide masses. The convergence of the two curves shows that the relative frequency of signature peptides increases with peptide mass.

The distribution of candidate signature transition set sizes across peptide masses is shown by the box plot underneath. Peptides without a unique product ion combination and peptides with signature transition sets consisting of more than three product ions were mostly found in the lower mass range between 600 - 1500 Da. Firstly, these peptides are short and therefore less likely to have a unique combination of product ions. Secondly, the background for these peptides is more densely populated. Peptides in the mass range between ∼1500 to ∼3500 Da were mostly identifiable by candidate signature transition sets consisting of one to three product ions.

The bar plot in the lower right corner shows the frequency of signature peptides for the respective signature transition set size. Around 10% of signature peptides (all in the mass range <1000 Da) did not have a unique product ion combination. The majority of peptides (∼75%) had unique combinations consisting of no more than three product ions.

Figure 4.9 B shows the proportion of protein sequences that emitted a signature peptide identifiable by a signature transition set of a given size. Signature peptides that feature a candidate signature transition set covered 99.8% of all proteins emitting a signature peptide when digested with the protease combination {Lys-C, Arg-C, V8E}. The large majority of covered proteins (98.6%) had at least one candidate signature transition set composed of no more than three product ion peaks.

4.3.2.2 Estimation of signature peptide detectability allowing for variable modification of methionine residues

Oxidation of methionine residues to methionine sulphoxide is a common protein modification occurring as an artifact of sample handling, e.g. purification or separation [412]. Methionine oxidation is a variable modification producing completely or partially oxidised peptides, resulting in a

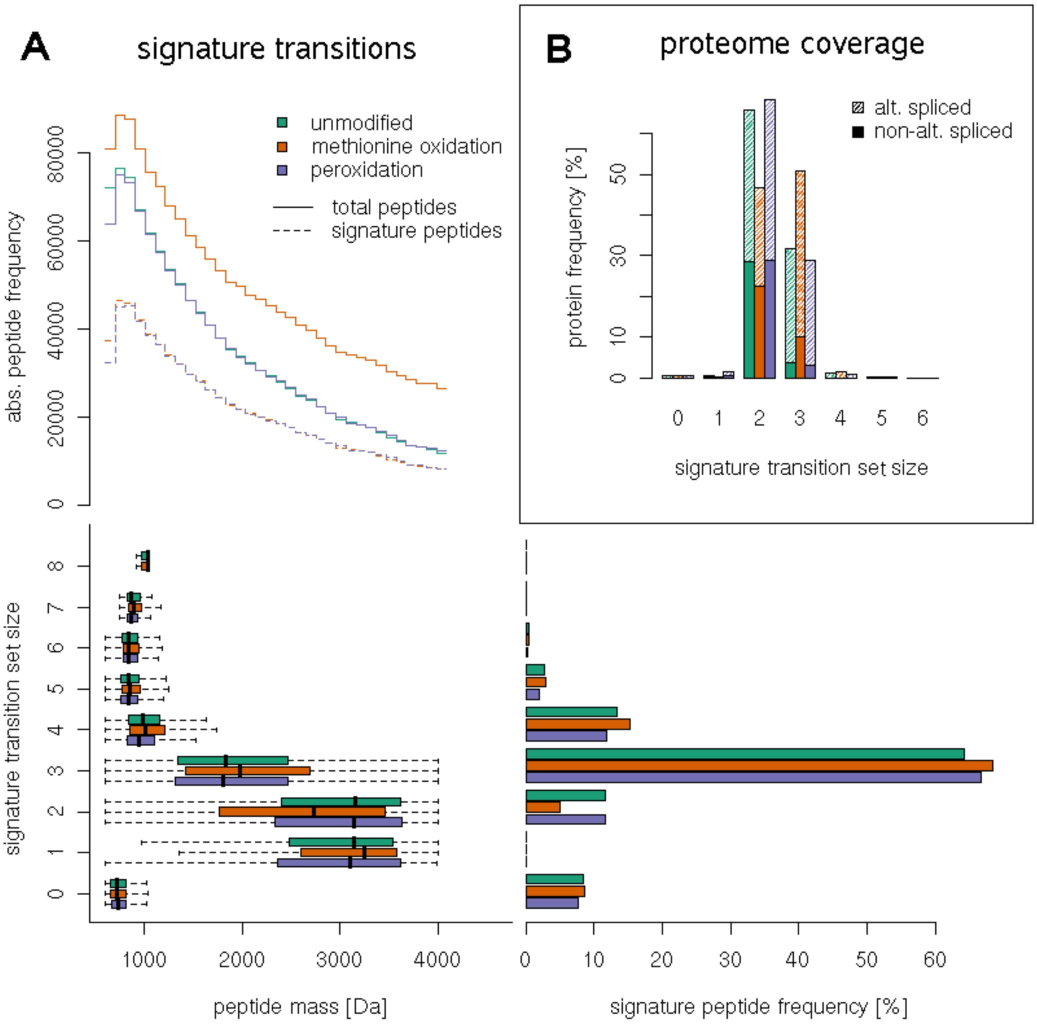


Figure 4.9: Signature transition set sizes and proteome coverage of signature transition sets for peptides generated by combinatorial proteolysis with Arg-C, Lys-C and V8D. A) The step plot shows the mass distribution of signature and background peptides, the box plot shows the distribution of signature transition set sizes across peptide masses, and the barplot shows the frequency of signature peptides identifiable by signature transitions of a given set size. B) Bar plot showing the proportion of protein sequences covered by signature transition sets of a given size. The shaded area represents protein products encoded by alternatively spliced genes.

series of peaks in the product ion spectrum that differ by 16 Da. Since this considerably increases the background against which a targeted peptide would have to be identified, targeted detectablility of signature peptides was assessed by searching for candidate signature transition sets in a peptide population generated by {Arg-C, Lys-C, V8E} exposed to variable methionine oxidatised.

The respective search results are shown in orange in figure 4.9. Due to the variable nature of the modification, the number of different peptide species to be taken into account in the search increased by 44%. The mass distribution plot shows that this negatively affected the relative frequency of signature peptides in general and neutralised the positive effect higher peptide masses have on the signature-to-background-peptide ratio. The result was a shift of signature transition frequencies towards higher set sizes: the number of peptides identifiable by two product ions decreasesd while the number of peptides that require three, four or five product ions for unique identification increased. The proportion of signature peptides without any unique combination stayed the same. On protein level this translated into a drop of proteins identifiable by signature transition sets of size two and an increase in proteins identifiable by signature transition sets consisting of three product ions.

4.3.2.3 The effect of performic acid oxidation on the detection of signature peptides

Performic acid oxidation is a classic chemical modification of proteins resulting in static oxidation of methionine, cysteine and tryptophane to $MetO_2$, $CysO_3$ and $TrpO_2$ [413]. To check if "force-oxidising" the protein sample, and thereby fixing residue modification, could antagonise the negative effect of variable methionine oxidation on signature peptide detectability, the search was repeated with statically oxidised methione, cysteine, and tryptophane residues.

The results of the search are shown in purple in figure 4.9. They suggest that performic acid oxidation does not only counter the negative effect of methionine oxidation but that it can restore the detectibility of signature peptides even above the level estimated for the unmodified sample. This is the result of i) fixing the modifications and ii) expanding the peptide mass distribution by introducing additional *O* atoms, resulting in a higher number of unique precursor and product ion masses [414].

4.4 Discussion

4.4.1 Increasing the scope of targeted proteomics by combinatorial proteolysis

Combining MS identifications from complementary peptide populations generated by different proteases has been employed previously to increase sequence coverage [57, 410, 58]. Here the applicablility of combinatorial proteolysis to the disambiguation of protein identifications, a key feature of emerging targeted proteomics approaches, was investigated. To this end proteolytic peptides generated by parallel combinatorial proteolysis were analysed in the context of the annotated genome. Using the genome as reference space (rather than a non-redundandent protein sequence databases) to analyse proteomics data makes sense as i) the proteome as defined by the annotated genome represents an equitable tradeoff between comprehensiveness and stringency, and ii) it enables a genecentric view of proteomics data important, for example, for an evaluation of splice isoform resolution.

As discussed in the preceeding chapters, the latter is limited in current proteomics experiments. The *in silico* analysis presented here shows that even in an ideal scenario, where every peptide is equally likely to be observed, resolution on protein sequence level would be compromised in the context of current proteomics protocols due to the absence of unique tryptic peptides for a significant proportion of protein products. These unresolved proteins are mostly isoforms of alternatively spliced genes.

The results presented here demonstrate that this limitation can be overcome to a large extent by combinatorial proteolysis. Two to three proteases are sufficient to achieve a significant improvement in proteome coverage by signature peptides and a considerable increase in the coverage of splice isoforms. The advantage of combinatorial proteolysis over tryptic digests regarding isoform resolution is of particular significance in light of the importance of alternative splicing for the study of gene function in health and disease, as well as in the context of a systematic effort to map the human proteome.

The cost-benefit analysis identified a combination of three proteases that markedly increased proteome coverage by signature peptides at a relatively modest increase in complexity of the peptide mixture. Notably, this combination contains the proteases Arg-C and Lys-C, each enzyme having a restriction site in common with trypsin. Thus, the resulting peptides will contain a basic residue at the C-terminus advantaguous for sequencing by CID, a major reason for trypsin being the protease of choice in MS based protoemics. While retaining the beneficial CID properties of trypsin, cleaving proteins at arginine and lysine residues in two separate digests results in longer peptides which are more likely to have a unique sequence and thus higher selectivity. This is the reason why the Arg-C, Lys-C combination results in a higher signature peptide coverage of the proteome, despite having the same combined cleavage specificity as trypsin.

In addition to a higher absolute number of proteins emitting signature peptides, combinatorial proteolysis was also shown to increase the sequence coverage by signature peptides. This could potentially positively influence the detectability of signature peptides by generating signature peptides from different regions of a protein, thereby increasing the chance of producing peptides with physicochemical properties suitable for detection in a MS experiment. This has been shown previously in analyses of membrane proteins by combinatorial approaches [57, 410].

4.4.2 Signature peptide detectability

Pooling of complementary peptide mixtures increases the complexity of the analysed peptide sample in terms of the number of distinct peptide species in the mixture. However, the total number of peptides in the mixture should not be affected significantly as peptides are generated from the same amount of starting material used in a digest with a single protease. In fact, the diversification of peptide sequences should result in a greater spread of peptide masses and consequently elusion times of peptides in the mixture. This could potentially reduce the effect of overshadowing of less abundant peptides by highly abundant peptides.

In the context of targeted proteomics approaches, the complexity of the peptide mixture should play a less important role anyway, as peptides are selectively detected by SRM. SRM is up to two orders of magnitude more sensitive than regular 'full scan' modes of operation, and shows a linear response in a dynamic range of up to five orders of magnitude. This enables the detection of low-abundance peptides in highly complex mixtures [415].

The detectability of signature peptides by SRM was estimated in an exhaustive search of theoretical spectra for combinations of product ion masses unique to the spectrum of the targeted peptide. Based on the *in silico* results, even at a very conservative choice of mass tolerance of $\pm 1 m/z$ for precursor and fragement ions, most signature peptides should be detectable against the background of isobaric peptides generated by the proteases Lys-C, Arg-C and V8E by monitoring no more than two to three product ions.

The analysis was based on the theoretical peptide population resulting from a digest of the entire proteome. In light of the fact that only a subset of proteins and isoforms is actually expressed in any given tissue it is likely that the results represent a pessimistic estimate of signature peptide detectablity that can be considered as a baseline for experimental detection of signature peptides by SRM.

In this context it should also be noted that two important factors affecting background estimation were not taken into account in this analysis, namely sequence variation arising from coding single nucleotide polymorphisms (cSNPs) and post-translational protein modifications. However, consideration of the former does not pose a major additional challenge to the computational analysis, as the number of peptides affected by cSNPs is, at less than 10%, relatively low. The latter is to some extend negligible, at least in studies where post-translational modifications are not the focus, as the most common modifications such as phosphorylation and glycosylation are often removed prior to the MS analysis to reduce sample complexity.

The majority of signature peptides featuring unique transition sets fall in the mass range between ∼1500 - ∼3500 Da which is optimal for detection by MS. Again, this is a result of longer peptides having a higher number of possible product ion combinations and thus a higher likelihood of producing a unique combination. Signature transition sets of peptides in the mass range below 1500 Da mostly consisted of four to eight product ions, which for short peptides is a large part of or even the entire spectrum making these transitions less interesting for use in an SRM approach.

Finally, the results indicate that performic acid oxidation has an overall beneficial effect on the detectability of signature peptides. Firstly, "force-oxidising" methionine counters the negative effect on detectability caused by variable methionine oxidation and the increase in complexity of the background that comes along with it. Secondly, the oxidation leads to an expansion of peptide masses resulting in less overlap of precursor as well as product ion masses, which in turn leads to a decrease in the number of transitions required to identify a peptide unambiguously.

4.4.3 Conclusions and perspectives

The analyses presented here addressed two issues frequently neglected in current discussions of targeted proteomics approaches as a means to overcome the shortcomings of shotgun proteomics regarding selectivity and sensitivity: firstly, the scope of such approaches in the context of the analysed peptide population and secondly, the information content of the targeted peptide reporters in a given experimental context. Most current efforts are centered around empirical information obtained from global proteomics approaches, and focus on the prediction of detectable peptides. However, peptides detectable in untargeted experiments might not necessarily be the most well performing in a targeted approach nor be the most informative in the context of a given biological question.

The strategy presented here inverses the process of identifying suitable peptides for the development of targeted proteomics experiments by first maximising the number of candidate reporters available for targeted identification through combinatorial proteolysis followed by an estimation of their information content in the context of a specific peptide population. The resulting pool of informative peptides can then be mined to select those peptides most suitable to address a specific question. Such an approach avoids disregarding highly selective and potentially highly responsive reporter peptides on the basis of predictions obtained from models trained on data generated by non-targeted, global proteomics approaches.

As currently available algorithms are trained on physicochemical properties of observed or highly observed peptides, biases in the predictions are likely to creep in for several reasons. Firstly, not only physicochemical properties but also abundance plays an important role in the detection of peptides by conventional proteomics approaches. Thus, predictions are likely to be biased towards physicochemical properties of highly abundant proteins which make up a large part of the data generated by shotgun proteomics. Secondly, peptides from proteins "disadvantaged" in mass spectrometry because of their physicochemical properties such as transmembrane proteins, are under-represented in the training data, resulting in a reduced sensitivity of prediction algorithms for such protein classes. Thirdly, an algorithm trained on peptide data generated in a different organism is biased because of differences in the amino acid composition of proteins between species. Another limitation arising from the exclusive use of empirical information to identify suitable SRM targets is the dominance of tryptic peptide data, restricting predictions to experiments using trypsin to generate peptides.

A potential drawback of the proposed reverse approach, on the other hand, is the significant increase in the number of candidate signature peptides that need to be tested for the development of transitions. However, in practice the number of peptides suitable to be reporter peptides can be limited in targeted experiments by the selectivity requirements. For instance, in case of alternatively spliced proteins only a relatively small set of peptides will be available to discriminate between specific isoforms. Furthermore, the evaluation of peptide information content carried out here does not only take into account uniqueness of the peptide sequence but also uniqueness of the transitions monitored by SRM. Information on precursor and peptide redundancy allows a ranking of peptides based on the level of specificity of precursor and product ion m/z values, prioritising those peptides with a lower background of isobaric precursor peptides and highly descriminative product ion m/z combinations.

None of the work published to date accounts for the implications of precursor and product ion redundancy on SRM specificity, an issue recently raised in a View-point article in the journal *Proteomics* [416]. Effectively the analysis presented here is an evaluation of the minimal requirements for unique peptide identification by combinations of multiple transitions as proposed by the authors. It demonstrates that specific transition combinations are available for the majority of signature pepides, even in complex mixtures generated by multiple proteases. As the estimation is based on a signature peptide having at least one unique transition set, and most peptides have more than one such set, chances of the existence of a specific product ion combination observable by MS are increased. Furthermore, the relatively small size of minimal transitions sets leaves scope for an increase in specificity by addition of further product ions if required.

The information on transition specificity made available by this study is a valuable resource for the prediction of SRM transitions in the future. In addition to the already discussed prioritisation based on discriminatory performance, candidate transitions could be selected for experimental validation based on their “flyability”, i.e. the ionization and/or detection efficiencies of precursor and product ions. One way of scoring the flyability of candidate signature peptides is the application of empirical rules based on the amino acid composition of the precursor peptide an approach taken by the SRM transition design tool MRMaid [404] for example. Amino acids with basic, positively charged amino acids like lysine, arginine, and histidine are positively weighted as they favour fragmentation by CID [417]. Residues that are subject to post-translational modification like tyrosin, serine, threonine (phosphorylation), cystein (carboxyamidomethylation) or methionine (oxidation) on the other hand negatively influence the suitability of signature peptide candidates as the mass shifts resulting from modification increase the complexity of precursor and product ion spectra [418]. Furthermore, it has been demonstrated that modified peptides show different ionisation and/or detection properties from their unmodified counterparts [419].

In addition to the weighting based on amino acid content product ion combinations could be prioritised by integration of the results with predictions of peptide fragmentation patterns [420, 123, 122, 421] and fragment intensities [422].

Alternatively, observable product ion combinations can be determined experimentally using synthetic peptides. Ideally such efforts would eventually lead to the construction of publicly available libraries of experimentally validated signature transition sets for use in targeted proteomics experiments.

Chapter 5

Conclusion

5.1 Summary

Two decades of mass spectrometry based protein analysis and nearly a decade of post-genome proteomics have demonstrated that mass spectrometry is a powerful and indispensable tool for the study of protein function at the systems level. Incremental methodological, technological and informatics advancements have led to significants improvements in shotgun mass spectrometry. Increasing coordination of worldwide proteomics efforts through organisations like the Human Proteome Organisation (HUPO) enables the systematic large-scale collaborative studies required to tackle the challenges involved in studying proteins on a genome-wide scale.

In the first chapter we saw that collaboration adds value to large-scale proteomics by increasing the scope of a study beyond what is achievable by single laboratories or groups. However, the results also emphasised the importance of study design and clearly defined endpoints to maximise the gain of collaborative efforts with regard to what can be learned about the technologies applied and, even more importantly, to better study the biology of the system of interest. Another recurring theme in large-scale proteomics studies, including the one discussed here, concerns problems associated with the biases and heterogeneity introduced by technological and methodological limitations.

The different mass spectrometry approaches applied in the study discussed in the first chapter were indeed able to identify a significant number of proteins critical for the system under study, albeit at varying levels of performance. Nevertheless, biases in the observed proteins resulted in the under-representation or absence of important functional classes as well as a limited resolution of isoforms. Chapter two showed that the trends identified on a study level translate into similar biases on a global level, and pointed out the extent of redundancy of mass spectrometry (MS) proteomics data. A clear conclusion from this investigation is the requirement for data exchange of the largely complementary results to unify information currently scattered over many proteomics resources. Such a strategy will increase the body of information that can be readily mined to improve both data analysis as well as the development of predictive methods in MS proteomics.

With the shortcomings of shotgun proteomics explored in detail in the first two chapters, the third chapter was dedicated to the targeted proteomics approaches that have increasingly captured the attention of the field. Two issues important in targeted proteomics workflows were addressed. The first of these was the availability of signature peptides to target a given protein. The limitations of tryptic peptide populations with regard to proteome coverage by signature peptides were demonstrated, in particular with respect to alternative splice isoforms. The potential of a strategy based on combinatorial proteolysis to overcome this constraint was analysed and found to be advantageous. The second issue derived from the possibility of having precursor and product ion redundancies. Since the ability to uniquely identify a peptide based on a limited number of transitions is critical to the selectivity of targeted proteomics experiments, the feasibility of this approach was assessed on a genome-wide scale. Even under the very conservative conditions chosen for the analysis, with constraints that underestimated the performance of current mass spectrometry technology, the results indicate that targeted proteomics strategies based on signature peptide reporters should be applicable to the majority of protein sequences. The information on unique transition sets for each signature peptide obtained through the analysis should make a contribution towards the ability to monitor signature peptide reporters in complex samples.

5.2 Outlook

With a Human Proteome Project (HPP) looming on the horizon, it is timely for the proteomics community to critically review its strategies and learn from the successes and failures of previous systematic efforts. Although a massively large-scale project like the HPP might benefit from a centralised organisational structure in which a few dedicated centres are leading the way, proteomics as a whole can only thrive on a mix of competition and collaboration rather than despotism. Regardless of the actual structure of the project, a high level of coordination and clearly defined goals will be central to the success of any future large-scale projects in proteomics. Indeed, the problems that proteomics is struggling with today are in large part the result of the proteome being a diffusely defined, moving target.

Although questioned by some, the benefits of a systematic mapping of human proteins are readily apparent. Indeed, although a systematic project like the HPP might at first seem uninspiring and of little relevance to basic research, it will be able to lay important foundations for future hypothesis-driven proteomics research. Additionally, it will facilitate integration within the field as well as with other 'omics' disciplines.

On the technical side it is becoming increasingly clear that some limitations of global proteomics approaches such as the high degree of redundancy, limited reproducibility and the difficulty of resolving low abundance proteins will be hard to overcome with shotgun approaches. While high-fidelity technologies like Fourier transform (FT)-ion cyclotron resonance (ICR) MS have greatly improved the accuracy of mass spectrometry, they cannot overcome the limitations relating to the complexity and dynamic range of the analysed samples. Furthermore, these technologies are expensive and currently only accessible to a privileged few.

Therefore, in the absence of any alternative technology for efficient protein identification comparable in performance to mass spectrometry, the trend towards targeted MS approaches is likely to continue. Not only are these approaches based on more affordable technologies, but they also solve, to a large extent, many of the limitations of global approaches. In particular in hypothesis-driven research, which is often focused on a limited, defined subset of the proteome, targeted workflows will probably enjoy increasing popularity. Furthermore, the increasing availability of protein-specific affinity reagents will allow for complementation of MS approaches by early stage technologies like protein arrays.

Finally, in light of the fast emerging targeted strategies, the role of bioinformatics in MS proteomics is undergoing a paradigm shift. Present-day proteomics bioinformatics is a tool primarily applied after the experiment has been conducted, to make sense of the results. In targeted approaches, however, bioinformatics will play a much more important role in up-front study design to identify suitable targets. In doing so it will be vital to effectively leverage information from genome sequences as well as from empirical data.

Although annotating the human proteome may seem a daunting task, rapid advancements in the available technologies, improved coordination of global proteomics efforts and new experimental strategies, may still allow the goal to be achieved earlier than anticipated.

Bibliography

[1] Smith AD, Datta DA S Pand Bender (2000) Oxford Dictionary of Biochemistry and Molecular Biology. New York: Oxford University Press.

[2] No authors listed (1995) 2D Electrophoresis: From Protein Maps to Genomes. Proceedings of the International Meeting. Siena, Italy, September 5-7, 1994. Electrophoresis 16:1077–1322.

[3] Wilkins M, Sanchez J, Gooley A, Appel R, Humphery-Smith I, et al. (1996) Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol Genet Eng Rev 13:19–50.

[4] Shevchenko A, Jensen O, Podtelejnikov A, Sagliocco F, Wilm M, et al. (1996) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. Proc Natl Acad Sci USA 93:14440–14445.

[5] Stein L (2001) Genome annotation: from sequence to biology. Nat Rev Genet 2:493–503.

[6] Lander E, Linton L, Birren B, Nusbaum C, Zody M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921.

[7] Venter J, Adams M, Myers E, Li P, Mural R, et al. (2001) The sequence of the human genome. Science 291:1304–1351.

[8] International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945.

[9] Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, et al. (2004) The Ensembl automatic gene annotation system. Genome Res 14:942–950.

[10] Hubbard T, Barker D, Birney E, Cameron G, Chen Y, et al. (2002) The Ensembl genome database project. Nucleic Acids Res 30:38–41.

[11] Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. Nucleic Acids Res 37:D690–697.

[12] Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, et al. (2008) The vertebrate genome annotation (Vega) database. Nucleic Acids Res 36:D753–760.

[13] Brett D, Pospisil H, Valcrcel J, Reich J, Bork P (2002) Alternative splicing and genome complexity. Nat Genet 30:29–30.

[14] Mironov AA, Fickett JW, Gelfand MS (1999) Frequent alternative splicing of human genes. Genome Res 9:1288–1293.

[15] Byrd MP, Zamora M, Lloyd RE (2002) Generation of multiple isoforms of eukaryotic translation initiation factor 4GI by use of alternate translation initiation codons. Mol Cell Biol 22:4499–4511.

[16] Kochetov AV (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. Bioessays 30:683–691.

[17] Humphery-Smith I (2004) A human proteome project with a beginning and an end. Proteomics 4:2519–2521.

[18] Van Regenmortel MH (2004) Reductionism and complexity in molecular biology. Scientists now have the tools to unravel biological and overcome the limitations of reductionism. EMBO Rep 5:1016–1020.

[19] Westerhoff HV, Palsson BO (2004) The evolution of molecular biology into systems biology. Nat Biotechnol 22:1249–1252.

[20] Janeway C, Travers P, Walport M, Capra J (1996) Immunobiology: the immune system in health and disease. Current Biology London.

[21] Larsson K, Wester K, Nilsson P, Uhlén M, Hober S, et al. (2006) Multiplexed PrEST immunization for high-throughput affinity proteomics. J Immunol Methods 315:110–120.

[22] Uhlen M, Ponten F (2005) Antibody-based proteomics for human tissue profiling. Mol Cell Proteomics 4:384–393.

[23] Andersson A, Stromberg S, Backvall H, Kampf C, Uhlen M, et al. (2006) Analysis of protein expression in cell microarrays: a tool for antibody-based proteomics. J Histochem Cytochem 54:1413–1423.

[24] Pehr E (1950) Method for determination of the amino acid sequence in peptides. Acta Chemica Scandinavica 4:283–293.

[25] Edman P, Begg G (1967) A protein sequenator. Eur J Biochem 1:80–91.

[26] Laursen R (1971) Solid-phase Edman degradation. An automatic peptide sequencer. Eur J Biochem 20:89–102.

[27] Niall H (1973) Automated Edman degradation: the protein sequenator. Meth Enzymol 27:942–1010.

[28] De Hoffmann E, Stroobant V (2007) Mass spectrometry: principles and applications. Wiley-Interscience.

[29] Aston F (1919) A positive ray spectrograph. Taylor & Francis.

[30] Yamashita M, Fenn JB (1984) Negative ion production with the electrospray ion source. The Journal of Physical Chemistry 88:4671–4675.

[31] Karas M, Bachmann D, Hillenkamp F (1985) Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. Analytical Chemistry 57:2935–2939.

[32] Tanaka K, Waki H, Y I, S A, Y Y, et al. (1988) Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Communications in Mass Spectrometry 2:151–153.

[33] Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science 246:64–71.

[34] Hillenkamp F, Karas M, Beavis RC, Chait BT (1991) Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. Anal Chem 63:1193A–1203A.

[35] Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, et al. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proc Natl Acad Sci USA 90:5011–5015.

[36] Ferguson PL, Smith RD (2003) Proteome analysis by mass spectrometry. Annu Rev Biophys Biomol Struct 32:399–424.

[37] Rabilloud T (2002) Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. Proteomics 2:3–10.

[38] Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc Natl Acad Sci USA 97:9390–9395.

[39] Martin SE, Shabanowitz J, Hunt DF, Marto JA (2000) Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. Anal Chem 72:4266–4274.

[40] Washburn MP, Wolters D, Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19:242–247.

[41] Malmström, J and Lee, H and Aebersold, R (2007) Advances in proteomic workflows for systems biology. Curr Opin Biotechnol 18:378–384.

[42] Florens L, Korfali N, Schirmer EC (2008) Subcellular fractionation and proteomics of nuclear envelopes. Methods Mol Biol 432:117–137.

[43] Milosevic J, Bulau P, Mortz E, Eickelberg O (2009) Subcellular fractionation of TGF-beta1-stimulated lung epithelial cells: a novel proteomic approach for identifying signaling intermediates. Proteomics 9:1230–1240.

[44] Lübke T, Lobel P, Sleat DE (2009) Proteomics of the lysosome. Biochim Biophys Acta 1793:625–635.

[45] Zhang J, Liem DA, Mueller M, Wang Y, Zong C, et al. (2008) Altered proteome biology of cardiac mitochondria under stress conditions. J Proteome Res 7:2204–2214.

[46] Gilchrist A, Au CE, Hiding J, Bell AW, Fernandez-Rodriguez J, et al. (2006) Quantitative proteomics analysis of the secretory pathway. Cell 127:1265–1281.

[47] Gevaert K, Van Damme P, Ghesquire B, Impens F, Martens L, et al. (2007) A la carte proteomics with an emphasis on gel-free techniques. Proteomics 7:2698–2718.

[48] Fountoulakis M, Juranville JF, Jiang L, Avila D, Röder D, et al. (2004) Depletion of the high-abundance plasma proteins. Amino Acids 27:249–259.

[49] Greenough C, Jenkins RE, Kitteringham NR, Pirmohamed M, Park BK, et al. (2004) A method for the rapid depletion of albumin and immunoglobulin from human plasma. Proteomics 4:3107–3111.

[50] Righetti PG, Boschetti E, Lomas L, Citterio A (2006) Protein Equalizer Technology : the quest for a ″democratic proteome″. Proteomics 6:3980–3992.

[51] Steen H, Mann M (2004) The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol 5:699–711.

[52] Nesvizhskii AI, Aebersold R (2005) Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 4:1419–1440.

[53] Kuster B, Schirle M, Mallick P, Aebersold R (2005) Scoring proteomes with proteotypic peptide probes. Nat Rev Mol Cell Biol 6:577–583.

[54] Olsen J, Ong S, Mann M (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. Mol Cell Proteomics 3:608–614.

[55] Picotti P, Aebersold R, Domon B (2007) The implications of proteolytic background for shotgun proteomics. Mol Cell Proteomics 6:1589–1598.

[56] Michalski WP, Shiell BJ (1999) Strategies for analysis of electrophoretically separated proteins and peptides. Analytica Chimica Acta 383:27 – 46.

[57] Fischer F, Poetsch A (2006) Protein cleavage strategies for an improved analysis of the membrane proteome. Proteome Sci 4:2.

[58] Schlosser A, Vanselow J, Kramer A (2005) Mapping of phosphorylation sites by a multi-protease approach with specific phosphopeptide enrichment and NanoLC-MS/MS analysis. Anal Chem 77:5243–5250.

[59] Mohammed S, Lorenzen K, Kerkhoven R, van Breukelen B, Vannini A, et al. (2008) Multiplexed proteomics mapping of yeast RNA polymerase II and III allows near-complete sequence coverage and reveals several novel phosphorylation sites. Anal Chem 80:3584–3592.

[60] Nordhoff E, Lehrach H (2007) Identification and characterization of DNA-binding proteins by mass spectrometry. Adv Biochem Eng Biotechnol 104:111–195.

[61] Chi A, Huttenhower C, Geer LY, Coon JJ, Syka JE, et al. (2007) Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. Proc Natl Acad Sci USA 104:2193–2198.

[62] Borchers CH, Thapar R, Petrotchenko EV, Torres MP, Speir JP, et al. (2006) Combined top-down and bottom-up proteomics identifies a phosphorylation site in stem-loop-binding proteins that contributes to high-affinity RNA binding. Proc Natl Acad Sci USA 103:3094–3099.

[63] Harsha H, H M, Pandey A (2008) Quantitative proteomics using stable isotope labeling with amino acids in cell culture. Nature Protocols 3:505–516.

[64] Han J, Schey KL (2004) Proteolysis and mass spectrometric analysis of an integral membrane: aquaporin 0. J Proteome Res 3:807–812.

[65] Zhang L, Eugeni EE, Parthun MR, Freitas MA (2003) Identification of novel histone post-translational modifications by peptide mass fingerprinting. Chromosoma 112:77–86.

[66] Berg J, Tymoczko J, Stryer L (2002) Biochemistry, 5th edition. WH Freeman and Company New York.

[67] Schirle M, Heurtier MA, Kuster B (2003) Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics 2:1297–1305.

[68] Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, et al. (1999) Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol 17:676–682.

[69] Motoyama A, Venable JD, Ruse CI, Yates JR (2006) Automated ultra-high-pressure multidimensional protein identification technology (UHP-MudPIT) for improved peptide identification of proteomic samples. Anal Chem 78:5109–5118.

[70] Linden J, Lawhead C (1975) Liquid chromatography of saccharides. J chromatogr 105:125–133.

[71] Moritz RL, Ji H, Schütz F, Connolly LM, Kapp EA, et al. (2004) A proteome strategy for fractionating proteins and peptides using continuous free-flow electrophoresis coupled off-line to reversed-phase high-performance liquid chromatography. Anal Chem 76:4811–4824.

[72] Boersema PJ, Divecha N, Heck AJ, Mohammed S (2007) Evaluation and optimization of ZIC-HILIC-RP as an alternative MudPIT strategy. J Proteome Res 6:937–946.

[73] Boersema PJ, Mohammed S, Heck AJ (2008) Hydrophilic interaction liquid chromatography (HILIC) in proteomics. Anal Bioanal Chem 391:151–159.

[74] Yi EC, Marelli M, Lee H, Purvine SO, Aebersold R, et al. (2002) Approaching complete peroxisome characterization by gas-phase fractionation. Electrophoresis 23:3205–3216.

[75] Kennedy J, Yi EC (2008) Use of gas-phase fractionation to increase protein identifications : application to the peroxisome. Methods Mol Biol 432:217–228.

[76] Ingvar Eidhammer I, Flikka K, Martens L, S M (2007) Computational Methods for Mass Spectrometry Proteomics. Wiley-Interscience.

[77] Stump M, Fleming R, Gong W, Jaber A, Jones J, et al. (2002) Matrix-assisted Laser Desorption Mass Spectrometry. Applied Spectroscopy Reviews 37:275–303.

[78] Zhang X, Narcisse DA, Murray KK (2004) On-line single droplet deposition for MALDI mass spectrometry. J Am Soc Mass Spectrom 15:1471–1477.

[79] Domon B, Aebersold R (2006) Mass spectrometry and protein analysis. Science 312:212–217.

[80] Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422:198–207.

[81] Hutchens T, Yip T (1993) New desorption strategies for the mass spectrometric analysis of macromolecules. Rapid Communications in Mass Spectrometry 7:576–580.

[82] Paul W, Steinwedel H (1953) Ein neues massenspektrometer ohne magnetfeld. Zeitschrift Naturforschung Teil A 8.

[83] Hager J, et al. (2002) A new linear ion trap mass spectrometer. Rapid Communications in Mass Spectrometry 16:512–526.

[84] Schwartz J, Senko M, Syka J (2002) A two-dimensional quadrupole ion trap mass spectrometer. Journal of the American Society for Mass Spectrometry 13:659–669.

[85] Hu Q, Noll RJ, Li H, Makarov A, Hardman M, et al. (2005) The Orbitrap: a new mass spectrometer. J Mass Spectrom 40:430–443.

[86] Makarov A (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem 72:1156–1162.

[87] Makarov A, Denisov E, Lange O, Horning S (2006) Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer. J Am Soc Mass Spectrom 17:977–982.

[88] Stephens WE (1946) A pulsed mass spectrometer with time dispersion. Phys Rev 69:691–692.

[89] Brown RS, Lennon JJ (1995) Mass resolution improvement by incorporation of pulsed ion extraction in a matrix-assisted laser desorption/ionization linear time-of-flight mass spectrometer. Anal Chem 67:1998–2003.

[90] Mamyrin B, Karataev V, Shmikk D, Zagulin V (1973) The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution. Sov Phys JETP 37:45–48.

[91] Paul W, Reinhard H, Von Zahn U (1958) Das elektrische Massenfilter als Massenspektrometer und Isotopentrenner. Zeitschrift für Physik 152:143–182.

[92] Stahl-Zeng J, Lange V, Ossola R, Eckhardt K, Krek W, et al. (2007) High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. Mol Cell Proteomics 6:1809–1817.

[93] Unwin RD, Griffiths JR, Leverentz MK, Grallert A, Hagan IM, et al. (2005) Multiple reaction monitoring to identify sites of protein phosphorylation with high sensitivity. Mol Cell Proteomics 4:1134–1144.

[94] Marshall AG, Hendrickson CL, Jackson GS (1998) Fourier transform ion cyclotron resonance mass spectrometry: a primer. Mass Spectrom Rev 17:1–35.

[95] Pappin D, Hojrup P, Bleasby A (1993) Rapid identification of proteins by peptide-mass fingerprinting. Curr Biol 3:327–332.

[96] Yates JR, Speicher S, Griffin PR, Hunkapiller T (1993) Peptide mass maps: a highly informative approach to protein identification. Anal Biochem 214:397–408.

[97] Loboda AV, Krutchinsky AN, Bromirski M, Ens W, Standing KG (2000) A tandem quadrupole/time-of-flight mass spectrometer with a matrix-assisted laser desorption/ionization source: design and performance. Rapid Commun Mass Spectrom 14:1047–1057.

[98] Shevchenko A, Loboda A, Shevchenko A, Ens W, Standing KG (2000) MALDI quadrupole time-of-flight mass spectrometry: a powerful tool for proteomic research. Anal Chem 72:2132–2141.

[99] Shevchenko A, Chernushevich I, Ens W, Standing KG, Thomson B, et al. (1997) Rapid 'de novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. Rapid Commun Mass Spectrom 11:1015–1024.

[100] Le Blanc JC, Hager JW, Ilisiu AM, Hunter C, Zhong F, et al. (2003) Unique scanning capabilities of a new hybrid linear ion trap mass spectrometer (Q TRAP) used for high sensitivity proteomics applications. Proteomics 3:859–869.

[101] Syka JE, Marto JA, Bai DL, Horning S, Senko MW, et al. (2004) Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. J Proteome Res 3:621–626.

[102] Peterman S, Dufresne C, Horning S (2005) The use of a hybrid linear trap/FT-ICR mass spectrometer for on-line high resolution/high mass accuracy bottom-up sequencing. Journal of Biomolecular Techniques: JBT 16:112.

[103] Mann M, Kelleher NL (2008) Precision proteomics: the case for high resolution and high mass accuracy. Proc Natl Acad Sci USA 105:18132–18138.

[104] Cleveland D, Fischer S, Kirschner M, Laemmli U (1977) Peptide mapping by limited proteolysis in sodium dodecyl sulfate and analysis by gel electrophoresis. J Biol Chem 252:1102–1106.

[105] Mann M, Hojrup P, Roepstorff P (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol Mass Spectrom 22:338–345.

[106] Rasmussen HH, Mortz E, Mann M, Roepstorff P, Celis JE (1994) Identification of transformation sensitive proteins recorded in human two-dimensional gel protein databases by mass spectrometric peptide mapping alone and in combination with microsequencing. Electrophoresis 15:406–416.

[107] Conrads TP, Anderson GA, Veenstra TD, Pasa-Tolic L, Smith RD (2000) Utility of accurate mass tags for proteome-wide protein identification. Anal Chem 72:3349–3354.

[108] Smith RD, Anderson GA, Lipton MS, Pasa-Tolic L, Shen Y, et al. (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. Proteomics 2:513–523.

[109] Lipton MS, Pasa-Tolic′ L, Anderson GA, Anderson DJ, Auberry DL, et al. (2002) Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. Proc Natl Acad Sci USA 99:11049–11054.

[110] White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, et al. (1999) Genome sequence of the radioresistant bacterium Deinococcus radiodurans R1. Science 286:1571–1577.

[111] Norbeck AD, Monroe ME, Adkins JN, Anderson KK, Daly DS, et al. (2005) The utility of accurate mass and LC elution time information in the analysis of complex proteomes. J Am Soc Mass Spectrom 16:1239–1249.

[112] Shinoda K, Sugimoto M, Tomita M, Ishihama Y (2008) Informatics for peptide retention properties in proteomic LC-MS. Proteomics 8:787–798.

[113] Mann M, Wilm M (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 66:4390–4399.

[114] Yates JR, Eng JK, McCormack AL, Schieltz D (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem 67:1426–1436.

[115] Fischer B, Roth V, Roos F, Grossmann J, Baginsky S, et al. (2005) NovoHMM: a hidden Markov model for de novo peptide sequencing. Anal Chem 77:7265–7273.

[116] Pevtsov S, Fedulova I, Mirzaei H, Buck C, Zhang X (2006) Performance evaluation of existing de novo sequencing algorithms. J Proteome Res 5:3018–3028.

[117] Pitzer E, Masselot A, Colinge J (2007) Assessing peptide de novo sequencing algorithms performance on large and diverse data sets. Proteomics 7:3051–3054.

[118] McHugh L, Arthur JW (2008) Computational methods for protein identification from mass spectrometry data. PLoS Comput Biol 4:e12.

[119] Wysocki VH, Resing KA, Zhang Q, Cheng G (2005) Mass spectrometry of peptides and proteins. Methods 35:211–222.

[120] Roepstorff P, Fohlman J (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom 11:601.

[121] Biemann K (1999) Nomenclature for peptide fragment ions. Methods in Enzymology 193:886–888.

[122] Zhang Z (2004) Prediction of low-energy collision-induced dissociation spectra of peptides. Anal Chem 76:3908–3922.

[123] Klammer AA, Reynolds SM, Bilmes JA, MacCoss MJ, Noble WS (2008) Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. Bioinformatics 24:i348–356.

[124] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3567.

[125] Fenyö D, Beavis RC (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem 75:768–774.

[126] Bafna V, Edwards N (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics 17 Suppl 1:13–21.

[127] Colinge J, Masselot A, Giron M, Dessingy T, Magnin J (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 3:1454–1463.

[128] Käll L, Storey JD, MacCoss MJ, Noble WS (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res 7:29–34.

[129] Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74:5383–5392.

[130] Resing KA, Meyer-Arendt K, Mendoza AM, Aveline-Wolf LD, Jonscher KR, et al. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. Anal Chem 76:3556–3568.

[131] Association of Biomolecular Resource Facilities, ABRF '05 (2005) Scaffold: a program to probabilistically combine results from multiple MS/MS database search engine. Abstract P87-T.

[132] Hamacher M, Apweiler R, Arnold G, Becker A, Blüggel M, et al. (2006) HUPO Brain Proteome Project: summary of the pilot phase and introduction of a comprehensive data reprocessing strategy. Proteomics 6:4890–4898.

[133] Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, et al. (2009) Protein identification false discovery rates for very large proteomics datasets generated by tandem mass spectrometry. Mol Cell Proteomics .

[134] Price TS, Lucitt MB, Wu W, Austin DJ, Pizarro A, et al. (2007) EBP, a program for protein identification using multiple tandem mass spectrometry datasets. Mol Cell Proteomics 6:527–536.

[135] Nesvizhskii A, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75:4646–4658.

[136] Sadygov RG, Liu H, Yates JR (2004) Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. Anal Chem 76:1664–1671.

[137] Sadygov RG, Yates JR (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal Chem 75:3792–3798.

[138] Weatherly DB, Atwood JA, Minning TA, Cavola C, Tarleton RL, et al. (2005) A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. Mol Cell Proteomics 4:762–772.

[139] Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 6:386–398.

[140] Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72:291–336.

[141] Tazi J, Bakkour N, Stamm S (2009) Alternative splicing and disease. Biochim Biophys Acta 1792:14–26.

[142] Brett D, Hanke J, Lehmann G, Haase S, Delbrück S, et al. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett 474:83–86.

[143] Modrek B, Resch A, Grasso C, Lee C (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 29:2850–2859.

[144] Eyras E, Caccamo M, Curwen V, Clamp M (2004) ESTGenes: alternative splicing from ESTs in Ensembl. Genome Res 14:976–987.

[145] Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302:2141–2144.

[146] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40:1413–1415.

[147] Tress ML, Bodenmiller B, Aebersold R, Valencia A (2008) Proteomics studies confirm the presence of alternative protein isoforms on a large scale. Genome Biol 9:R162.

[148] Tanner S, Shen Z, Ng J, Florea L, Guigo R, et al. (2007) Improving gene annotation using peptide mass spectrometry. Genome Res 17:231–239.

[149] Desiere F, Deutsch E, Nesvizhskii A, Mallick P, King N, et al. (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol 6:R9.

[150] Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, et al. (2006) Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biol 7:R35.

[151] Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD (2008) Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief Funct Genomic Proteomic 7:50–62.

[152] Nesvizhskii AI, Vitek O, Aebersold R (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 4:787–797.

[153] Zhou H, Watts JD, Aebersold R (2001) A systematic approach to the analysis of protein phosphorylation. Nat Biotechnol 19:375–378.

[154] Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, et al. (2002) Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. Nat Biotechnol 20:301–305.

[155] Knight ZA, Schilling B, Row RH, Kenski DM, Gibson BW, et al. (2003) Phosphospecific proteolysis for mapping sites of protein phosphorylation. Nat Biotechnol 21:1047–1054.

[156] Schaefer H, Chamrad DC, Marcus K, Reidegeld KA, Blüggel M, et al. (2005) Tryptic transpeptidation products observed in proteome analysis by liquid chromatography-tandem mass spectrometry. Proteomics 5:846–852.

[157] Grubb RL, Calvert VS, Wulkuhle JD, Paweletz CP, Linehan WM, et al. (2003) Signal pathway profiling of prostate cancer using reverse phase protein arrays. Proteomics 3:2142–2146.

[158] Gembitsky DS, Lawlor K, Jacovina A, Yaneva M, Tempst P (2004) A prototype antibody microarray platform to monitor changes in protein tyrosine phosphorylation. Mol Cell Proteomics 3:1102–1118.

[159] Perez OD, Nolan GP (2006) Phospho-proteomic immune analysis by flow cytometry: from mechanism to translational medicine at the single-cell level. Immunol Rev 210:208–228.

[160] Steen JA, Steen H, Georgi A, Parker K, Springer M, et al. (2008) Different phosphorylation states of the anaphase promoting complex in response to antimitotic drugs: a quantitative proteomic analysis. Proc Natl Acad Sci USA 105:6069–6074.

[161] Tao WA, Wollscheid B, O'Brien R, Eng JK, Li XJ, et al. (2005) Quantitative phosphoproteome analysis using a dendrimer conjugation chemistry and tandem mass spectrometry. Nat Methods 2:591–598.

[162] Cantin GT, Venable JD, Cociorva D, Yates JR (2006) Quantitative phosphoproteomic analysis of the tumor necrosis factor pathway. J Proteome Res 5:127–134.

[163] Blagoev B, Ong SE, Kratchmarova I, Mann M (2004) Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. Nat Biotechnol 22:1139–1145.

[164] López-Otín C, Overall CM (2002) Protease degradomics: a new challenge for proteomics. Nat Rev Mol Cell Biol 3:509–519.

[165] Puente XS, Snchez LM, Overall CM, Lpez-Otn C (2003) Human and mouse proteases: a comparative genomic approach. Nat Rev Genet 4:544–558.

[166] Neurath H (1999) Proteolytic enzymes, past and future. Proc Natl Acad Sci USA 96:10962–10963.

[167] Dean RA, Overall CM (2007) Proteomics discovery of metalloproteinase substrates in the cellular context by iTRAQ labeling reveals a diverse MMP-2 substrate degradome. Mol Cell Proteomics 6:611–623.

[168] Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, et al. (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. Nat Biotechnol 21:566–569.

[169] Hwang IK, Park SM, Kim SY, Lee ST (2004) A proteomic approach to identify substrates of matrix metalloproteinase-14 in human plasma. Biochim Biophys Acta 1702:79–87.

[170] Overall CM, Tam EM, Kappelhoff R, Connor A, Ewart T, et al. (2004) Protease degradomics: mass spectrometry discovery of protease substrates and the CLIP-CHIP, a dedicated DNA microarray of all human proteases and inhibitors. Biol Chem 385:493–504.

[171] Lee AY, Park BC, Jang M, Cho S, Lee DH, et al. (2004) Identification of caspase-3 degradome by two-dimensional gel electrophoresis and matrix-assisted laser desorption/ionization-time of flight analysis. Proteomics 4:3429–3436.

[172] Van Damme P, Martens L, Van Damme J, Hugelier K, Staes A, et al. (2005) Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis. Nat Methods 2:771–777.

[173] Apweiler R, Hermjakob H, Sharon N (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochim Biophys Acta 1473:4–8.

[174] Arnold JN, Wormald MR, Sim RB, Rudd PM, Dwek RA (2007) The impact of glycosylation on the biological function and structure of human immunoglobulins. Annu Rev Immunol 25:21–50.

[175] Zhao YY, Takahashi M, Gu JG, Miyoshi E, Matsumoto A, et al. (2008) Functional roles of N-glycans in cell signaling and cell adhesion in cancer. Cancer Sci 99:1304–1310.

[176] Dwek RA (1998) Biological importance of glycosylation. Dev Biol Stand 96:43–47.

[177] Molinari M (2007) N-glycan structure dictates extension of protein folding or onset of disposal. Nat Chem Biol 3:313–320.

[178] Gaynor EC, Emr SD (1997) COPI-independent anterograde transport: cargo-selective ER to Golgi protein transport in yeast COPI mutants. J Cell Biol 136:789–802.

[179] Dell A, Morris HR (2001) Glycoprotein structure determination by mass spectrometry. Science 291:2351–2356.

[180] Morelle W, Flahaut C, Michalski JC, Louvet A, Mathurin P, et al. (2006) Mass spectrometric approach for screening modifications of total serum N-glycome in human diseases: application to cirrhosis. Glycobiology 16:281–293.

[181] Ethier M, Saba JA, Spearman M, Krokhin O, Butler M, et al. (2003) Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. Rapid Commun Mass Spectrom 17:2713–2720.

[182] Manzi AE, Norgard-Sumnicht K, Argade S, Marth JD, van Halbeek H, et al. (2000) Exploring the glycan repertoire of genetically modified mice by isolation and profiling of the major glycan classes and nano-NMR analysis of glycan mixtures. Glycobiology 10:669–689.

[183] Rudd PM, Colominas C, Royle L, Murphy N, Hart E, et al. (2001) A high-performance liquid chromatography based strategy for rapid, sensitive sequencing of N-linked oligosaccharide modifications to proteins in sodium dodecyl sulphate polyacrylamide electrophoresis gel bands. Proteomics 1:285–294.

[184] Hashii N, Kawasaki N, Itoh S, Hyuga M, Kawanishi T, et al. (2005) Glycomic/glycoproteomic analysis by liquid chromatography/mass spectrometry: analysis of glycan structural alteration in cells. Proteomics 5:4665–4672.

[185] Blixt O, Head S, Mondala T, Scanlan C, Huflejt ME, et al. (2004) Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. Proc Natl Acad Sci USA 101:17033–17038.

[186] Song X, Xia B, Lasanajak Y, Smith DF, Cummings RD (2008) Quantifiable fluorescent glycan microarrays. Glycoconj J 25:15–25.

[187] Zheng T, Peelen D, Smith LM (2005) Lectin arrays for profiling cell surface carbohydrate expression. J Am Chem Soc 127:9982–9983.

[188] Kuno A, Uchiyama N, Koseki-Kuno S, Ebe Y, Takashima S, et al. (2005) Evanescent-field fluorescence-assisted lectin microarray: a new strategy for glycan profiling. Nat Methods 2:851–856.

[189] Uchiyama N, Kuno A, Tateno H, Kubo Y, Mizuno M, et al. (2008) Optimization of evanescent-field fluorescence-assisted lectin microarray for high-sensitivity detection of monovalent oligosaccharides and glycoproteins. Proteomics 8:3042–3050.

[190] Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402:47–52.

[191] Vidal M (2001) A biological atlas of functional maps. Cell 104:333–339.

[192] Young KH (1998) Yeast two-hybrid: so many interactions, (in) so little time.. Biol Reprod 58:302–311.

[193] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403:623–627.

[194] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98:4569–4574.

[195] Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. (2003) A protein interaction map of Drosophila melanogaster. Science 302:1727–1736.

[196] Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al. (2004) A map of the interactome network of the metazoan C. elegans. Science 303:540–543.

[197] Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147.

[198] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437:1173–1178.

[199] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122:957–968.

[200] Ewing RM, Chu P, Elisma F, Li H, Taylor P, et al. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol 3:89.

[201] Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, et al. (2004) A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. Nat Cell Biol 6:97–105.

[202] Falsone SF, Gesslbauer B, Tirk F, Piccinini AM, Kungl AJ (2005) A proteomic snapshot of the human heat shock protein 90 interactome. FEBS Lett 579:6350–6354.

[203] Blagoev B, Kratchmarova I, Ong SE, Nielsen M, Foster LJ, et al. (2003) A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. Nat Biotechnol 21:315–318.

[204] Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–1016.

[205] Guda C, Subramaniam S (2005) pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes. Bioinformatics 21:3963–3969.

[206] Cokol M, Nair R, Rost B (2000) Finding nuclear localization signals. EMBO Rep 1:411–415.

[207] Garg A, Bhasin M, Raghava GP (2005) Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem 280:14427–14432.

[208] Eisenhaber B, Bork P, Eisenhaber F (1999) Prediction of potential GPI-modification sites in proprotein sequences. J Mol Biol 292:741–758.

[209] Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580.

[210] Tusndy GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17:849–850.

[211] Simpson JC, Wellenreuther R, Poustka A, Pepperkok R, Wiemann S (2000) Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. EMBO Rep 1:287–292.

[212] Liebel U, Starkuviene V, Erfle H, Simpson JC, Poustka A, et al. (2003) A microscope-based screening platform for large-scale functional protein analysis in intact cells. FEBS Lett 554:394–398.

[213] Wiemann S, Arlt D, Huber W, Wellenreuther R, Schleeger S, et al. (2004) From ORFeome to biology: a functional genomics pipeline. Genome Res 14:2136–2144.

[214] Bannasch D, Mehrle A, Glatting KH, Pepperkok R, Poustka A, et al. (2004) LIFEdb: a database for functional genomics experiments integrating information from external sources, and serving as a sample tracking system. Nucleic Acids Res 32:D505–508.

[215] Yates JR, Gilchrist A, Howell KE, Bergeron JJ (2005) Proteomics of organelles and large cellular structures. Nat Rev Mol Cell Biol 6:702–714.

[216] Andersen JS, Lam YW, Leung AK, Ong SE, Lyon CE, et al. (2005) Nucleolar proteome dynamics. Nature 433:77–83.

[217] Andersen JS, Lyon CE, Fox AH, Leung AK, Lam YW, et al. (2002) Directed proteomic analysis of the human nucleolus. Curr Biol 12:1–11.

[218] Schirmer EC, Florens L, Guan T, Yates JR, Gerace L (2003) Nuclear membrane proteins with potential disease links found by subtractive proteomics. Science 301:1380–1382.

[219] Breuza L, Halbeisen R, Jenö P, Otte S, Barlowe C, et al. (2004) Proteomics of endoplasmic reticulum-Golgi intermediate compartment (ERGIC) membranes from brefeldin A-treated HepG2 cells identifies ERGIC-32, a new cycling protein that interacts with human Erv46. J Biol Chem 279:47242–47253.

[220] Bell AW, Ward MA, Blackstock WP, Freeman HN, Choudhary JS, et al. (2001) Proteomics characterization of abundant Golgi membrane proteins. J Biol Chem 276:5152–5165.

[221] Pisitkun T, Shen RF, Knepper MA (2004) Identification and proteomic profiling of exosomes in human urine. Proc Natl Acad Sci USA 101:13368–13373.

[222] Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, et al. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. Nature 426:570–574.

[223] Sauer G, Körner R, Hanisch A, Ries A, Nigg EA, et al. (2005) Proteome analysis of the human mitotic spindle. Mol Cell Proteomics 4:35–43.

[224] Taylor SW, Fahy E, Zhang B, Glenn GM, Warnock DE, et al. (2003) Characterization of the human heart mitochondrial proteome. Nat Biotechnol 21:281–286.

[225] Dunkley TP, Watson R, Griffin JL, Dupree P, Lilley KS (2004) Localization of organelle proteins by isotope tagging (LOPIT). Mol Cell Proteomics 3:1128–1134.

[226] Sadowski PG, Dunkley TP, Shadforth IP, Dupree P, Bessant C, et al. (2006) Quantitative proteomic approach to study subcellular localization of membrane proteins. Nat Protoc 1:1778–1789.

[227] Foster LJ, de Hoog CL, Zhang Y, Zhang Y, Xie X, et al. (2006) A mammalian organelle map by protein correlation profiling. Cell 125:187–199.

[228] Hall SL, Hester S, Griffin JL, Lilley KS, Jackson AP (2009) The organelle proteome of the DT40 Lymphocyte cell line. Mol Cell Proteomics .

[229] Su A, Wiltshire T, Batalov S, Lapp H, Ching K, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101:6062–6067.

[230] Uhlén M, Björling E, Agaton C, Szigyarto CA, Amini B, et al. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. Mol Cell Proteomics 4:1920–1932.

[231] Nilsson P, Paavilainen L, Larsson K, Odling J, Sundberg M, et al. (2005) Towards a human proteome atlas: high-throughput generation of mono-specific antibodies for tissue profiling. Proteomics 5:4327–4337.

[232] Warford A (2004) Tissue microarrays: fast-tracking protein expression at the cellular level. Expert Rev Proteomics 1:283–292.

[233] Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, et al. (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. Proteomics 5:3226–3245.

[234] States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, et al. (2006) Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. Nat Biotechnol 24:333–338.

[235] Xie H, Rhodus NL, Griffin RJ, Carlis JV, Griffin TJ (2005) A catalogue of human saliva proteins identified by free flow electrophoresis-based peptide separation and tandem mass spectrometry. Mol Cell Proteomics 4:1826–1830.

[236] Pan S, Zhu D, Quinn J, Peskind E, Montine T, et al. (2007) A combined dataset of human cerebrospinal fluid proteins identified by multi-dimensional chromatography and tandem mass spectrometry. Proteomics 7:469–473.

[237] Ji J, Chakraborty A, Geng M, Zhang X, Amini A, et al. (2000) Strategy for qualitative and quantitative analysis in proteomics based on signature peptides. J Chromatogr B Biomed Sci Appl 745:197–210.

[238] Taylor J, Anderson N, Scandora A, Willard K, Anderson N (1982) Design and implementation of a prototype Human Protein Index. Clin Chem 28:861–866.

[239] Abbott A (2001) Workshop prepares ground for human proteome project. Nature 413:763.

[240] Hanash S, Celis JE (2002) The Human Proteome Organization: a mission to advance proteome knowledge. Mol Cell Proteomics 1:413–414.

[241] Uhlen M (2008) A new era for proteomics research? Genome Biol 9:325.

[242] Cottingham K (2008) HUPO's Human Proteome Project: the next big thing? Journal of Proteome Research 7:2192–2192.

[243] Anderson NL, Anderson NG, Pearson TW, Borchers CH, Paulovich AG, et al. (2009) A human proteome detection and quantitation project: hPDQ. Mol Cell Proteomics .

[244] Krijgsveld J, Whetton A, Lee B, Lemischka I, Oh S, et al. (2008) Proteome biology of stem cells: a new joint HUPO and ISSCR initiative. Mol Cell Proteomics 7:204–205.

[245] Hamacher M, Meyer H (2005) HUPO Brain Proteome Project: aims and needs in proteomics. Expert Rev Proteomics 2:1–3.

[246] Omenn G (2004) International collaboration in clinical chemistry and laboratory medicine: the Human Proteome Organization (HUPO) Plasma Proteome Project. Clin Chem Lab Med 42:1–2.

[247] Cottingham K (2007) Proteomics projects: Hupo cardiovascular initiative comes into its own. Journal of Proteome Research 6:1242–1242.

[248] Cottingham K (2008) Proteomics projects: Hkupp becomes a full-fledged initiative of its own. Journal of Proteome Research 7:484–484.

[249] Paik Y (2006) Disease Biomarker Discovery in Korea. Proteomics 6:1091–1093.

[250] Haab B, Paulovich A, Anderson N, Clark A, Downing G, et al. (2006) A reagent resource to identify proteins and peptides of interest for the cancer community: a workshop report. Mol Cell Proteomics 5:1996–2007.

[251] Orchard S, Hermjakob H, Apweiler R (2003) The proteomics standards initiative. Proteomics 3:1374–1376.

[252] Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, et al. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. Nat Biotechnol 21:247–254.

[253] Deutsch E (2008) mzML: a single, unifying data format for mass spectrometer output. Proteomics 8:2776–2777.

[254] Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, et al. (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. Nat Biotechnol 22:177–183.

[255] Prince JT, Carlson MW, Wang R, Lu P, Marcotte EM (2004) The need for a public proteomics repository. Nat Biotechnol 22:471–472.

[256] Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, et al. (2005) PRIDE: the proteomics identifications database. Proteomics 5:3537–3545.

[257] Craig R, Cortens J, Beavis R (2004) Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 3:1234–1242.

[258] Mathivanan S, Ahmed M, Ahn NG, Alexandre H, Amanchy R, et al. (2008) Human Proteinpedia enables sharing of human protein data. Nat Biotechnol 26:164–167.

[259] Mead JA, Shadforth IP, Bessant C (2007) Public proteomic MS repositories and pipelines: available tools and biological applications. Proteomics 7:2769–2786.

[260] Mead JA, Bianco L, Bessant C (2009) Recent developments in public proteomic MS repositories and pipelines. Proteomics 9:861–881.

[261] Keller A, Eng J, Zhang N, Li XJ, Aebersold R (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol 1:2005.0017.

[262] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410.

[263] Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, et al. (2004) UniProt archive. Bioinformatics 20:3236–3237.

[264] Jones P, Ct RG, Cho SY, Klie S, Martens L, et al. (2008) PRIDE: new developments and new datasets. Nucleic Acids Res 36:D878–883.

[265] Smedley D, Haider S, Ballester B, Holland R, London D, et al. (2009) BioMart–biological queries made easy. BMC Genomics 10:22.

[266] Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, et al. (2004) Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res 32:497–501.

[267] Hermjakob H, Apweiler R (2006) The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. Expert Rev Proteomics 3:1–3.

[268] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32:D115–119.

[269] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365–370.

[270] Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, et al. (2003) The Protein Information Resource. Nucleic Acids Res 31:345–347.

[271] Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, et al. (2005) The EMBL Nucleotide Sequence Database. Nucleic Acids Res 33:29–33.

[272] O'Donovan C, Apweiler R, Bairoch A (2001) The human proteomics initiative (HPI). Trends Biotechnol 19:178–181.

[273] Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The UCSC Genome Browser Database. Nucleic Acids Res 31:51–54.

[274] http://www.ncbi.nlm.nih.gov/projects/mapview.

[275] Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2009) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 37:5–15.

[276] Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, et al. (2004) The International Protein Index: an integrated database for proteomics experiments. Proteomics 4:1985–1988.

[277] Ashburner M, Ball C, Blake J, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

[278] Goeman JJ, Mansmann U (2008) Multiple testing on the directed acyclic graph of gene ontology. Bioinformatics 24:537–544.

[279] Ct RG, Jones P, Apweiler R, Hermjakob H (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinformatics 7:97.

[280] Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, et al. (2003) eVOC: a controlled vocabulary for unifying gene expression data. Genome Res 13:1222–1230.

[281] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2002) InterPro: an integrated documentation resource for protein families, domains and functional sites. Brief Bioinformatics 3:225–235.

[282] Boguski M, Lowe T, Tolstoshev C (1993) dbEST–database for "expressed sequence tags". Nat Genet 4:332–333.

[283] Su A, Cooke M, Ching K, Hakak Y, Walker J, et al. (2002) Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci USA 99:4465–4470.

[284] Hanash S (2004) HUPO initiatives relevant to clinical proteomics. Mol Cell Proteomics 3:298–301.

[285] No authors listed (2005) Proteomics' new order. Nature 437:169–170.

[286] HUPO HPO (2005) Establishment of HUPO: Mission and Objectives. Published Online .

[287] Calabria A, Shusta E (2006) Blood-brain barrier genomics and proteomics: elucidating phenotype, identifying disease targets and enabling brain drug delivery. Drug Discov Today 11:792–799.

[288] Papassotiropoulos A, Fountoulakis M, Dunckley T, Stephan D, Reiman E (2006) Genetics, transcriptomics, and proteomics of Alzheimer's disease. J Clin Psychiatry 67:652–670.

[289] Enard W, Khaitovich P, Klose J, Zöllner S, Heissig F, et al. (2002) Intra- and interspecific variation in primate gene expression patterns. Science 296:340–343.

[290] Schmidt O, Schulenborg T, Meyer H, Marcus K, Hamacher M (2005) How proteomics reveals potential biomarkers in brain diseases. Expert Rev Proteomics 2:901–913.

[291] Johnson M, Yu L, Conrads T, Kinoshita Y, Uo T, et al. (2005) The proteomics of neurodegeneration. Am J Pharmacogenomics 5:259–270.

[292] Berven F, Flikka K, Berle M, Vedeler C, Ulvik R (2006) Proteomic-based biomarker discovery with emphasis on cerebrospinal fluid and multiple sclerosis. Curr Pharm Biotechnol 7:147–158.

[293] Schulenborg T, Schmidt O, van Hall A, Meyer H, Hamacher M, et al. (2006) Proteomics in neurodegeneration–disease driven approaches. J Neural Transm 113:1055–1073.

[294] Lovestone S, Güntert A, Hye A, Lynham S, Thambisetty M, et al. (2007) Proteomics of Alzheimer's disease: understanding mechanisms and seeking biomarkers. Expert Rev Proteomics 4:227–238.

[295] Andrade E, Krueger D, Nairn A (2007) Recent advances in neuroproteomics. Curr Opin Mol Ther 9:270–281.

[296] Lewczuk P, Esselmann H, Otto M, Maler J, Henkel A, et al. (2004) Neurochemical diagnosis of Alzheimer's dementia by CSF Abeta42, Abeta42/Abeta40 ratio and total tau. Neurobiol Aging 25:273–281.

[297] Otto M, Lewczuk P, Wiltfang J (2008) Neurochemical approaches of cerebrospinal fluid diagnostics in neurodegenerative diseases. Methods 44:289–298.

[298] Romeo M, Espina V, Lowenthal M, Espina B, Petricoin E, et al. (2005) CSF proteome: a protein repository for potential biomarker identification. Expert Rev Proteomics 2:57–70.

[299] Milner B (1975) Psychological aspects of focal epilepsy and its neurosurgical management. Adv Neurol 8:299–321.

[300] Victor M, Adams R (2005) Principles of Neurology - 8th Edition. New York: McGraw-Hill.

[301] Berg A (2008) The natural history of mesial temporal lobe epilepsy. Curr Opin Neurol 21:173–178.

[302] Liu X, Yang J, Chen L, Zhang Y, Yang M, et al. (2008) Comparative proteomics and correlated signaling network of rat hippocampus in the pilocarpine model of temporal lobe epilepsy. Proteomics 8:582–603.

[303] Wang Q, Woltjer R, Cimino P, Pan C, Montine K, et al. (2005) Proteomic analysis of neurofibrillary tangles in Alzheimer disease identifies GAPDH as a detergent-insoluble paired helical filament tau binding protein. FASEB J 19:869–871.

[304] Zouambia M, Fischer D, Hobo B, De Vos R, Hol E, et al. (2008) Proteasome subunit proteins and neuropathology in tauopathies and synucleinopathies: Consequences for proteomic analyses. Proteomics 8:1221–1236.

[305] Porchet R, Probst A, Bouras C, Dráberová E, Dráber P, et al. (2003) Analysis of glial acidic fibrillary protein in the human entorhinal cortex during aging and in Alzheimer's disease. Proteomics 3:1476–1485.

[306] Martins-de Souza D, Gattaz WF, Schmitt A, Rewerts C, Marangoni S, et al. (2009) Alterations in oligodendrocyte proteins, calcium homeostasis and new potential markers in schizophrenia anterior temporal lobe are revealed by shotgun proteome analysis. J Neural Transm 116:275–289.

[307] Habeck M (2003) Brain proteome project launched. Nature Medicine 9:631–631.

[308] Abbott A (2003) Brain protein project enlists mice in 'dry run'. Nature 425:110.

[309] Dumont D, Noben J, Verhaert P, Stinissen P, Robben J (2006) Gel-free analysis of the human brain proteome: application of liquid chromatography and mass spectrometry on biopsy and autopsy samples. Proteomics 6:4967–4977.

[310] Park Y, Kim J, Kwon K, Lee S, Kim Y, et al. (2006) Profiling human brain proteome by multi-dimensional separations coupled with MS. Proteomics 6:4978–4986.

[311] He S, Wang Q, He J, Pu H, Yang W, et al. (2006) Proteomic analysis and comparison of the biopsy and autopsy specimen of human brain temporal lobe. Proteomics 6:4987–4996.

[312] Stühler K, Pfeiffer K, Joppich C, Stephan C, Jung K, et al. (2006) Pilot study of the Human Proteome Organisation Brain Proteome Project: applying different 2-DE techniques to monitor proteomic changes during murine brain development. Proteomics 6:4899–4913.

[313] Föcking M, Boersema P, O'Donoghue N, Lubec G, Pennington S, et al. (2006) 2-D DIGE as a quantitative tool for investigating the HUPO Brain Proteome Project mouse series. Proteomics 6:4914–4931.

[314] Seefeldt I, Nebrich G, Römer I, Mao L, Klose J (2006) Evaluation of 2-DE protein patterns from pre- and postnatal stages of the mouse brain. Proteomics 6:4932–4939.

[315] Carrette O, Burkhard P, Hochstrasser D, Sanchez J (2006) Age-related proteome analysis of the mouse brain: a 2-DE study. Proteomics 6:4940–4949.

[316] Fröhlich T, Helmstetter D, Zobawa M, Crecelius A, Arzberger T, et al. (2006) Analysis of the HUPO Brain Proteome reference samples using 2-D DIGE and 2-D LC-MS/MS. Proteomics 6:4950–4966.

[317] Stephan C, Reidegeld K, Hamacher M, van Hall A, Marcus K, et al. (2006) Automated reprocessing pipeline for searching heterogeneous mass spectrometric data of the HUPO Brain Proteome Project pilot phase. Proteomics 6:5015–5029.

[318] Kar A, Kuo D, He R, Zhou J, Wu J (2005) Tau alternative splicing and frontotemporal dementia. Alzheimer Dis Assoc Disord 19 Suppl 1:29–36.

[319] Andreadis A (2006) Misregulation of tau alternative splicing in neurodegeneration and dementia. Prog Mol Subcell Biol 44:89–107.

[320] Licatalosi D, Darnell R (2006) Splicing regulation in neurologic disease. Neuron 52:93–101.

[321] Wu C, Apweiler R, Bairoch A, Natale D, Barker W, et al. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34:D187–191.

[322] Hubbard T, Aken B, Beal K, Ballester B, Caccamo M, et al. (2007) Ensembl 2007. Nucleic Acids Res 35:D610–617.

[323] Bendtsen J, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795.

[324] Sonnhammer E, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6:175–182.

[325] Mulder N, Apweiler R, Attwood T, Bairoch A, Bateman A, et al. (2007) New developments in the InterPro database. Nucleic Acids Res 35:D224–228.

[326] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, et al. (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–120.

[327] Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, et al. (2004) EnsMart: a generic system for fast and flexible access to biological data. Genome Res 14:160–169.

[328] Camon E, Magrane M, Barrell D, Lee V, Dimmer E, et al. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 32:D262–266.

[329] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, et al. (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 33:D428–432.

[330] Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37:D619–622.

[331] Hamosh A, Scott A, Amberger J, Bocchini C, McKusick V (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33:D514–517.

[332] Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. J Comput And Graph Stat 5:299–314.

[333] Wu Z, Irizarry R, Gentleman R, Martinez-Murillo F, Spencer F (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of the American Statistical Association 99:909–917.

[334] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc, Ser B 57:289–300.

[335] Benjamini Y, Kenigsberg E, Reiner A, D Y (2008) FDR adjustments of Microarray Experiments. Technical report, Department of Statistics and O.R., Tel Aviv Universityd. http://www.bioconductor.org/packages/2.3/bioc/vignettes/fdrame/inst/doc/fdrame.pdf.

[336] Povey S, Lovering R, Bruford E, Wright M, Lush M, et al. (2001) The HUGO Gene Nomenclature Committee (HGNC). Hum Genet 109:678–680.

[337] Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, et al. (2008) The HGNC Database in 2008: a resource for the human genome. Nucleic Acids Res 36:D445–448.

[338] Swiss Institute of Bioinformatics, European Bioinformatics Institute, Protein Information Resource (2009) UniProt Knowledge Base User Manual. http://www.expasy.org/sprot/userman.html#CC_line.

[339] Mallick P, Schirle M, Chen SS, Flory MR, Lee H, et al. (2007) Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 25:125–131.

[340] Birney E, Stamatoyannopoulos J, Dutta A, Guigó R, Gingeras T, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816.

[341] Pruess M, Fleischmann W, Kanapin A, Karavidopoulou Y, Kersey P, et al. (2003) The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. Nucleic Acids Res 31:414–417.

[342] Pasqualato S, Renault L, Cherfils J (2002) Arf, Arl, Arp and Sar proteins: a family of GTP-binding proteins with a structural device for 'front-back' communication. EMBO Rep 3:1035–1041.

[343] Schimmöller F, Simon I, Pfeffer S (1998) Rab GTPases, directors of vesicle docking. J Biol Chem 273:22161–22164.

[344] Raichle M, Gusnard D (2002) Appraising the brain's energy budget. Proc Natl Acad Sci USA 99:10237–10239.

[345] Nakakura E, Watkins D, Schuebel K, Sriuranpong V, Borges M, et al. (2001) Mammalian Scratch: a neural-specific Snail family transcriptional repressor. Proc Natl Acad Sci USA 98:4010–4015.

[346] Brunelli S, Silva Casey E, Bell D, Harland R, Lovell-Badge R (2003) Expression of Sox3 throughout the developing central nervous system is dependent on the combined action of discrete, evolutionarily conserved regulatory elements. Genesis 36:12–24.

[347] Söllner T, Bennett M, Whiteheart S, Scheller R, Rothman J (1993) A protein assembly-disassembly pathway in vitro that may correspond to sequential steps of synaptic vesicle docking, activation, and fusion. Cell 75:409–418.

[348] Schiavo G, Stenbeck G, Rothman J, Söllner T (1997) Binding of the synaptic vesicle v-SNARE, synaptotagmin, to the plasma membrane t-SNARE, SNAP-25, can explain docked vesicles at neurotoxin-treated synapses. Proc Natl Acad Sci USA 94:997–1001.

[349] Görg A, Weiss W, Dunn M (2004) Current two-dimensional electrophoresis technology for proteomics. Proteomics 4:3665–3685.

[350] Brodsky F, Chen C, Knuehl C, Towler M, Wakeham D (2001) Biological basket weaving: formation and function of clathrin-coated vesicles. Annu Rev Cell Dev Biol 17:517–568.

[351] Besnard F, Brenner M, Nakatani Y, Chao R, Purohit H, et al. (1991) Multiple interacting sites regulate astrocyte-specific transcription of the human gene for glial fibrillary acidic protein. J Biol Chem 266:18877–18883.

[352] Klie S, Martens L, Vizcaíno JA, Côté R, Jones P, et al. (2008) Analyzing large-scale proteomics projects with latent semantic indexing. J Proteome Res 7:182–191.

[353] Merrihew G, Davis C, Ewing B, Williams G, Käll L, et al. (2008) Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. Genome Res 18:1660–1669.

[354] Choudhary J, Grant S (2004) Proteomics in postgenomic neuroscience: the end of the beginning. Nat Neurosci 7:440–445.

[355] Côté RG, Jones P, Martens L, Kerrien S, Reisinger F, et al. (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. BMC Bioinformatics 8:401.

[356] Kahlem P, Birney E (2007) ENFIN a network to enhance integrative systems biology. Ann N Y Acad Sci 1115:23–31.

[357] GuptaRoy B, Beckingham K, Griffith LC (1996) Functional diversity of alternatively spliced isoforms of Drosophila Ca2+/calmodulin-dependent protein kinase II. A role for the variable domain in activation. J Biol Chem 271:19846–19851.

[358] Goulet I, Gauvin G, Boisvenue S, Cote J (2007) Alternative splicing yields protein arginine methyltransferase 1 isoforms with distinct activity, substrate specificity, and subcellular localization. J Biol Chem 282:33009–33021.

[359] Fagnani M, Barash Y, Ip JY, Misquitta C, Pan Q, et al. (2007) Functional coordination of alternative splicing in the mammalian central nervous system. Genome Biol 8:R108.

[360] Yeo G, Holste D, Kreiman G, Burge CB (2004) Variation in alternative splicing across human tissues. Genome Biol 5:R74.

[361] Bairoch A, Bougueleret L, Altairac S, Amendolia V, Auchincloss A, et al. (2009) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res 37:D169–174.

[362] UniProt Consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36:D190–195.

[363] Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, et al. (2008) Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science 320:938–941.

[364] de Souza GA, Malen H, Softeland T, Saelensminde G, Prasad S, et al. (2008) High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using Mycobacterium tuberculosis as an example. BMC Genomics 9:316.

[365] Buza TJ, McCarthy FM, Burgess SC (2007) Experimental-confirmation and functional-annotation of predicted proteins in the chicken genome. BMC Genomics 8:425.

[366] Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, et al. (2006) The PeptideAtlas project. Nucleic Acids Res 34:D655–658.

[367] Gnad F, Oroshi M, Birney E, Mann M (2009) MAPU 2.0: high-accuracy proteomes mapped to genomes. Nucleic Acids Res 37:D902–906.

[368] Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, et al. (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res 16:123–131.

[369] Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, et al. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128:1231–1245.

[370] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837.

[371] Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, et al. (2006) Ensembl 2006. Nucleic Acids Res 34:D556–561.

[372] Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2008) Ensembl 2008. Nucleic Acids Res 36:D707–714.

[373] Pearson H (2008) Biologists initiate plan to map human proteome. Nature 452:920–921.

[374] Hochstrasser D (2008) Should the Human Proteome Project Be Gene- or Protein-centric? J Proteome Res 7:5071–5071.

[375] Kandasamy K, Keerthikumar S, Goel R, Mathivanan S, Patankar N, et al. (2009) Human Proteinpedia: a unified discovery resource for proteomics research. Nucleic Acids Res 37:D773–781.

[376] Craig R, Cortens JP, Beavis RC (2005) The use of proteotypic peptide libraries for protein identification. Rapid Commun Mass Spectrom 19:1844–1850.

[377] Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448.

[378] Martens L, Vandekerckhove J, Gevaert K (2005) DBToolkit: processing protein databases for peptide-centric proteomics. Bioinformatics 21:3584–3585.

[379] Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, et al. (2008) Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics 60:1–18.

[380] Deutsch EW, Lam H, Aebersold R (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep 9:429–434.

[381] Walsh GM, Lin S, Evans DM, Khosrovi-Eghbal A, Beavis RC, et al. (2009) Implementation of a data repository-driven approach for targeted proteomics experiments by multiple reaction monitoring. J Proteomics 72:838–852.

[382] Mueller M, Vizcaino J, Jones P, Cote R, Thorneycroft D, et al. (2008) Analysis of the experimental detection of central nervous system-related genes in human brain and cerebrospinal fluid datasets. Proteomics 8:1138–1148.

[383] Martens L, Mueller M, Stephan C, Hamacher M, Reidegeld K, et al. (2006) A comparison of the HUPO Brain Proteome Project pilot with other proteomics studies. Proteomics 6:5076–5086.

[384] Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, et al. (1999) Mining SNPs from EST databases. Genome Res 9:167–174.

[385] Yates J, Eng J, McCormack A (1995) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal Chem 67:3202–3210.

[386] Neubauer G, King A, Rappsilber J, Calvio C, Watson M, et al. (1998) Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. Nat Genet 20:46–50.

[387] Edwards NJ (2007) Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol 3:102.

[388] Schandorff S, Olsen JV, Bunkenborg J, Blagoev B, Zhang Y, et al. (2007) A mass spectrometry-friendly database for cSNP identification. Nat Methods 4:465–466.

[389] Anderson NL, Anderson NG (2002) The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics 1:845–867.

[390] Echan LA, Tang HY, Ali-Khan N, Lee K, Speicher DW (2005) Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma. Proteomics 5:3292–3303.

[391] Taussig MJ, Stoevesandt O, Borrebaeck CA, Bradbury AR, Cahill D, et al. (2007) ProteomeBinders: planning a European resource of affinity reagents for analysis of the human proteome. Nat Methods 4:13–17.

[392] Pontén F, Jirström K, Uhlen M (2008) The Human Protein Atlas–a tool for pathology. J Pathol 216:387–393.

[393] Josic D, Clifton JG (2007) Mammalian plasma membrane proteomics. Proteomics 7:3010–3029.

[394] Lu B, McClatchy DB, Kim JY, Yates JR (2008) Strategies for shotgun identification of integral membrane proteins by tandem mass spectrometry. Proteomics 8:3947–3955.

[395] Sandhu C, Hewel JA, Badis G, Talukder S, Liu J, et al. (2008) Evaluation of data-dependent versus targeted shotgun proteomic approaches for monitoring transcription factor expression in breast cancer. J Proteome Res 7:1529–1541.

[396] Yocum AK, Gratsch TE, Leff N, Strahler JR, Hunter CL, et al. (2008) Coupled global and targeted proteomics of human embryonic stem cells during induced differentiation. Mol Cell Proteomics 7:750–767.

[397] Anderson L, Hunter CL (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol Cell Proteomics 5:573–588.

[398] Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM (2007) Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. Proc Natl Acad Sci USA 104:5860–5865.

[399] Ahmed N, Thornalley PJ (2003) Quantitative screening of protein biomarkers of early glycation, advanced glycation, oxidation and nitrosation in cellular and extracellular proteins by tandem mass spectrometry multiple reaction monitoring. Biochem Soc Trans 31:1417–1422.

[400] Ciccimaro E, Hevko J, Blair IA (2006) Analysis of phosphorylation sites on focal adhesion kinase using nanospray liquid chromatography/multiple reaction monitoring mass spectrometry. Rapid Commun Mass Spectrom 20:3681–3692.

[401] Hülsmeier AJ, Paesold-Burda P, Hennet T (2007) N-glycosylation site occupancy in serum glycoproteins using multiple reaction monitoring liquid chromatography-mass spectrometry. Mol Cell Proteomics 6:2132–2138.

[402] DeSouza LV, Taylor AM, Li W, Minkoff MS, Romaschin AD, et al. (2008) Multiple reaction monitoring of mTRAQ-labeled peptides enables absolute quantification of endogenous levels of a potential can-

cer marker in cancerous and normal endometrial tissues. J Proteome Res 7:3525–3534.

[403] Janecki DJ, Bemis KG, Tegeler TJ, Sanghani PC, Zhai L, et al. (2007) A multiple reaction monitoring method for absolute quantification of the human liver alcohol dehydrogenase ADH1C1 isoenzyme. Anal Biochem 369:18–26.

[404] Mead JA, Bianco L, Ottone V, Barton C, Kay RG, et al. (2009) MRMaid, the web-based tool for designing multiple reaction monitoring (MRM) transitions. Mol Cell Proteomics 8:696–705.

[405] Martin DB, Holzman T, May D, Peterson A, Eastham A, et al. (2008) MRMer, an interactive open source and cross-platform system for data extraction and visualization of multiple reaction monitoring experiments. Mol Cell Proteomics 7:2270–2278.

[406] Tang H, Arnold RJ, Alves P, Xun Z, Clemmer DE, et al. (2006) A computational approach toward label-free protein quantification using predicted peptide detectability. Bioinformatics 22:e481–488.

[407] Fusaro VA, Mani DR, Mesirov JP, Carr SA (2009) Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nat Biotechnol 27:190–198.

[408] Molina H, Horn D, Tang N, Mathivanan S, Pandey A (2007) Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. Proc Natl Acad Sci USA 104:2199–2204.

[409] MacCoss M, McDonald W, Saraf A, Sadygov R, Clark J, et al. (2002) Shotgun identification of protein modifications from protein complexes and lens tissue. Proc Natl Acad Sci USA 99:7900–7905.

[410] Distler A, Kerner J, Peterman S, Hoppel C (2006) A targeted proteomic approach for the analysis of rat liver mitochondrial outer membrane proteins with extensive sequence coverage. Anal Biochem 356:18–29.

[411] Wu S, Kim J, Hancock W, Karger B (2005) Extended Range Proteomic Analysis (ERPA): a new and sensitive LC-MS platform for high sequence coverage of complex proteins with extensive post-translational modifications-comprehensive analysis of beta-casein

and epidermal growth factor receptor (EGFR). J Proteome Res 4:1155–1170.

[412] Lagerwerf FM, van de Weert M, Heerma W, Haverkamp J (1996) Identification of oxidized methionine in peptides. Rapid Commun Mass Spectrom 10:1905–1910.

[413] Hirs CHW (1967) Performic acid oxidation. In: Hirs CHW, editor, Enzyme Structure, Academic Press, volume 11 of *Methods in Enzymology*. pp. 197 – 199.

[414] Matthiesen R, Bauw G, Welinder KG (2004) Use of performic acid oxidation to expand the mass distribution of tryptic peptides. Anal Chem 76:6848–6852.

[415] Lange V, Picotti P, Domon B, Aebersold R (2008) Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol 4:222.

[416] Sherman J, McKay MJ, Ashman K, Molloy MP (2009) How specific is my SRM?: The issue of precursor and product ion redundancy. Proteomics 9:1120–1123.

[417] Papayannopoulos I (1995) The interpretation of collision-induced dissociation tandem mass spectra of peptides. Mass Spectrometry Reviews 14.

[418] Bunkenborg J, Matthiesen R (2007) Interpretation of collision-induced fragmentation tandem mass spectra of posttranslationally modified peptides. Methods Mol Biol 367:169–194.

[419] Steen H, Jebanathirajah JA, Springer M, Kirschner MW (2005) Stable isotope-free relative and absolute quantitation of protein phosphorylation stoichiometry by MS. Proc Natl Acad Sci USA 102:3948–3953.

[420] Zhang Z (2005) Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. Anal Chem 77:6364–6373.

[421] Arnold R, Jayasankar N, Aggarwal D, Tang H, Radivojac P (2006) A machine learning approach to predicting peptide fragmentation spectra. Pac Symp Biocomput :219–230.

[422] Frank AM (2009) Predicting intensity ranks of peptide fragment ions. J Proteome Res 8:2226–2240.

[423] Mueller M, Martens L, Apweiler R (2007) Annotating the human proteome: beyond establishing a parts list. Biochim Biophys Acta 1774:175–191.

[424] Mueller M, Martens L, Reidegeld K, Hamacher M, Stephan C, et al. (2006) Functional annotation of proteins identified in human brain during the HUPO Brain Proteome Project pilot study. Proteomics 6:5059–5075.

[425] Vizcaíno J, Mueller M, Hermjakob H, Martens L (2009) Charting online OMICS resources: A navigational chart for clinical researchers. Proteomics - Clinical Aapplications 3:18–29.

[426] Yan W, Apweiler R, Balgley B, Boontheung P, Bundy J, et al. (2009) Systematic comparison of the human saliva and plasma proteomes. Proteomics - Clinical Aapplications 3:116–134.

[427] Zhang J, Li X, Mueller M, Wang Y, Zong C, et al. (2008) Systematic characterization of the murine mitochondrial proteome using functionally validated cardiac mitochondria. Proteomics 8:1564–1575.

[428] Reidegeld K, Mueller M, Stephan C, Blüggel M, Hamacher M, et al. (2006) The power of cooperative investigation: summary and comparison of the HUPO Brain Proteome Project pilot study results. Proteomics 6:4997–5014.

[429] Stephan C, Hamacher M, Blüggel M, Körting G, Chamrad D, et al. (2005) 5th HUPO BPP Bioinformatics Meeting at the European Bioinformatics Institute in Hinxton, UK–Setting the analysis frame. Proteomics 5:3560–3562.

[430] Hamacher M, Stephan C, Blüggel M, Chamrad D, Körting G, et al. (2006) The HUPO Brain Proteome Project jamboree: centralised summary of the pilot studies. Proteomics 6:1719–1721.

Appendix A

Publications & presentations

A.1 Publications

1. Mueller M, Vizcano JA, Jones P, Ct R, Thorneycroft D, Apweiler R, Hermjakob H, Martens L (2008) Analysis of the experimental detection of central nervous system-related genes in human brain and cerebrospinal fluid datasets. Proteomics 8:1138-1148. [382]

2. Mueller M, Martens L, Apweiler R (2007) Annotating the human proteome: beyond establishing a parts list. Biochim Biophys Acta 1774:175-191. [423]

3. Mueller M, Martens L, Reidegeld K, Hamacher M, Stephan C, Blueggel M, Koerting G, Chamrad D, Scheer C, Marcus K, Meyer HE, Apweiler R (2006) Functional annotation of proteins identified in human brain during the HUPO Brain Proteome Project pilot study. Proteomics 6:5059-5075. [424]

4. Martens L, Mueller M, Stephan C, Hamacher M, Reidegeld K, Meyer HE, Blueggel M, Vandekerckhove J, Gevaert K, Apweiler R (2006) A comparison of the HUPO Brain Proteome Project pilot with other proteomics studies. Proteomics 6:5076-5086. [383]

5. Mueller M, Vizcano JA, Hermjakob H, Martens L (2009) Charting online OMICS resources: A navigational chart for clinical researchers. Proteomics - Clinical Aapplications 3:18-29. [425]

6. Yan W, Apweiler R, Balgley BM, Boontheung P, Bundy JL, Cargile BJ, Cole S, Fang X, Gonzalez-Begne M, Griffin TJ, Hagen F, Hu S, Wolinsky LE, Lee CS, Malamud D, Melvin JE, Menon R, Mueller M,

Qiao R, Rhodus NL, Sevinsky JR, States D, Stephenson JL Jr, Than S, Yates III JR, Yu W, Xie H, Xie Y, Omenn GS, Loo JA, Wong ST (2009) Systematic comparison of the human saliva and plasma proteomes. Proteomics - Clinical Aapplications 3:116-132. [426]

7. Zhang J, Liem D, Mueller M, Wang Y, Zong C, Deng N, Vondriska TM, Korge P, Drews O, Maclellan WR, Honda H, Weiss JN, Apweiler R, Ping P (2008) Altered proteome biology of cardiac mitochondria under stress conditions. J Proteome Res 7:2204-2214. [45]

8. Zhang J, Li X, Mueller M,Wang Y, Zong C, Deng N, Vondriska TM, Liem DA, Yang JI, Korge P, Honda H, Weiss JN, Apweiler R, Ping P (2008) Systematic characterization of the murine mitochondrial proteome using functionally validated cardiac mitochondria. Proteomics 8:1564-1575. [427]

9. Reidegeld K, Mueller M, Stephan C, Blueggel M, Hamacher M, Martens L, Koerting G, Chamrad DC, Parkinson D, Apweiler R, Meyer HE, Marcus K (2006) The power of cooperative investigation: summary and comparison of the HUPO Brain Proteome Project pilot study results. Proteomics 6:4997-5014. [428]

10. Stephan C, Reidegeld K, Hamacher M, van Hall A, Marcus K, Taylor C, Jones P, Mueller M, Apweiler R, Martens L, Koerting G, Chamrad DC, Thiele H, Blueggel M, Parkinson D, Binz PA, Lyall A, Meyer HE (2006) Automated reprocessing pipeline for searching heterogeneous mass spectrometric data of the HUPO Brain Proteome Project pilot phase. Proteomics 6:5015-5029. [317]

11. Stuehler K, Pfeiffer K, Joppich C, Stephan C, Jung K, Mueller M, Schmidt O, van Hall A, Hamacher M, Urfer W, Meyer HE, Marcus K (2006) Pilot study of the Human Proteome Organisation Brain Proteome Project: applying different 2-DE techniques to monitor proteomic changes during murine brain development. Proteomics 6:4899-4913. [312]

12. Stephan C, Hamacher M, Blueggel M, Koerting G, Chamrad D, Scheer C, Marcus K, Reidegeld KA, Lohaus C, Schaefer H, Martens L, Jones P, Mueller M, Auyeung K, Taylor C, Binz PA, Thiele H, Parkinson D, Meyer HE, Apweiler R (2005) 5th HUPO BPP Bioinformatics Meeting at the European Bioinformatics Institute in Hinxton, UK - Setting the analysis frame. Proteomics 5:3560-3562. [429]

13. Hamacher M, Stephan C, Blggel M, Chamrad D, Koerting G, Martens L, Mueller M, Hermjakob H, Parkinson D, Dowsey A, Reidegeld KA, Marcus K, Dunn MJ, Meyer HE, Apweiler R (2006) The HUPO Brain Proteome Project jamboree: centralised summary of the pilot studies. Proteomics 6:1719-1721. [430]

A.2 Conference & workshop presentations

1. Estimating the scope and selectivity of a targeted proteomics approach based on combinatorial proteolysis, Michael Mueller, Lennart Martens, Henning Hermjakob, Rolf Apweiler, 8th Siena Meeting: From Genome to Proteome, Siena, Italy, September 2008 (oral presentation)

2. Large-scale proteomics experiments in context, Michael Mueller, Henning Hermjakob, Rolf Apweiler, Human Proteome Organisation 5th Annual World Congress, Long Beach, USA, November 2006 (poster presentation)

3. Putting proteomics experiments into context, Michael Mueller, Rolf Apweiler, 3rd Joint Meeting British Society for Proteome Research & European Bioinformatics Institute, Cambridge, UK, July 2006 (oral presentation)

4. The HUPO Brain Proteome Project: Annotating the human brain proteome, Michael Mueller and Rolf Apweiler, 5th Human Proteome Organisation Brain Proteome Project workshop, Dublin, Ireland, February 2006 (oral presentation)