

Bioinformatics solutions for confident identification and targeted quantification of proteins using tandem mass spectrometry

Jennifer A. Cham (née Mead)

Engineering Doctorate (EngD) Thesis
Cranfield Health, Cranfield University
October 2009

Supervisors: Dr. Conrad Bessant, Dr. Stephen Regan
Sponsor: GlaxoSmithKline Plc.

**This thesis is submitted in partial fulfilment of the requirements for the degree of
engineering doctorate (EngD)**

**©Cranfield University, 2009. All rights reserved. No part of this publication may
be reproduced without the written permission of the copyright holder.**

Abstract

Proteins are the structural supports, signal messengers and molecular workhorses that underpin living processes in every cell. Understanding when and where proteins are expressed, and their structure and functions, is the realm of proteomics. Mass spectrometry (MS) is a powerful method for identifying and quantifying proteins, however, very large datasets are produced, so researchers rely on computational approaches to transform raw data into protein information. This project develops new bioinformatics solutions to support the next generation of proteomic MS research.

Part I introduces the state of the art in proteomic bioinformatics in industry and academia. The business history and funding mechanisms are examined to fill a notable gap in management research literature, and to explain events at the sponsor, GlaxoSmithKline. It reveals that public funding of proteomic science has yet to come to fruition and exclusively high-tech niche bioinformatics businesses can succeed in the current climate. Next, a comprehensive review of repositories for proteomic MS is performed, to locate and compile a summary of sources of datasets for research activities in this project, and as a novel summary for the community. Part II addresses the issue of false positive protein identifications produced by automated analysis with a proteomics pipeline. The work shows that by selecting a suitable decoy database design, a statistically significant improvement in identification accuracy can be made. Part III describes development of computational resources for selecting multiple reaction monitoring (MRM) assays for quantifying proteins using MS. A tool for transition design, MRMAid (pronounced 'mermaid'), and database of pre-published transitions, MRMAid-DB, are developed, saving practitioners time and leveraging existing resources for superior transition selection.

By improving the quality of identifications, and providing support for quantitative approaches, this project brings the field a small step closer to achieving the goal of systems biology.

Table of Contents

ABSTRACT	I
TABLE OF CONTENTS.....	III
TABLE OF FIGURES	IX
TABLE OF TABLES.....	XIII
ACKNOWLEDGEMENTS.....	XV
DEDICATION.....	XVII
ABBREVIATIONS (SCIENCE)	XIX
AMINO ACID SINGLE LETTER ABBREVIATIONS.....	XXII
ABBREVIATIONS (BUSINESS).....	XXIII
THESIS STRUCTURE SUMMARY.....	XXV
1 INTRODUCTION.....	1
1.1 PROTEOMICS IS THE LARGE-SCALE STUDY OF PROTEINS	3
1.1.1 <i>Proteomic mass spectrometry is an approach used for biomarker discovery ...</i>	7
1.1.2 <i>Proteomics is an ‘omics’ science that is needed for systems biology</i>	9
1.1.3 <i>Proteomic bioinformatics exploits computers for analysing large datasets from proteomic MS experiments.....</i>	11
1.2 THE PRACTICALITIES OF HIGH-THROUGHPUT PROTEOMIC MS AND PROTEOME BIOINFORMATICS	12
1.2.1 <i>Proteins are prepared for MS using 2D-PAGE or MuDPIT.....</i>	12
1.2.2 <i>Mass spectrometry of peptides can be performed in high-throughput.....</i>	20
1.2.3 <i>Proteins can be quantified using MS.....</i>	34
1.2.4 <i>An introduction to proteomic bioinformatics.....</i>	42
1.3 AIMS AND OBJECTIVES OF THIS THESIS.....	66
2 THE BUSINESS HISTORY OF PROTEOMIC BIOINFORMATICS (1985-2009)...	71
2.1 SUMMARY	73

2.2	INTRODUCTION	74
2.2.1	<i>The proteomic bioinformatics market is investigated using a business history approach</i>	74
2.2.2	<i>Management hypothesis and contribution to knowledge</i>	76
2.3	THE BUSINESS HISTORY OF PROTEOMIC BIOINFORMATICS.....	78
2.3.1	<i>The early innovators of proteomics and proteomic bioinformatics were publicly-funded</i>	78
2.3.2	<i>High-throughput analysis became possible by applying mass spectrometry to proteins</i>	82
2.3.3	<i>Privately-funded biotechs competed for control during the land-grab</i>	86
2.3.4	<i>Big Pharma invested heavily in proteomic bioinformatics infrastructure</i>	94
2.4	CURRENT STATUS OF THE PROTEOMIC BIOINFORMATICS MARKET	101
2.4.1	<i>Proteomic bioinformatics research is publicly funded again</i>	101
2.4.2	<i>Company case studies illustrate the current market for proteomic bioinformatics products and services</i>	103
2.5	DISCUSSION.....	124
2.5.1	<i>High-throughput proteomics is a typical 'hype cycle' technology</i>	124
2.6	PROVING THE MANAGEMENT HYPOTHESIS	137
2.7	RECOMMENDATIONS FOR INVESTORS AND FUNDING BODIES FOR PROTEOMIC BIOINFORMATICS	139
2.7.1	<i>Funding high-tech can be a springboard for growth or a futile cycle</i>	139
2.7.2	<i>Specialise for success: the contract research model is recommended</i>	146
2.8	CONCLUSION.....	150
3	REVIEW OF PUBLIC REPOSITORIES FOR PROTEOMICS	153
3.1	SUMMARY	155
3.2	INTRODUCTION	156
3.2.1	<i>Definition of a public proteomics repository</i>	156
3.2.2	<i>Benefits of public repositories</i>	156
3.3	AN OVERVIEW OF THE PUBLIC REPOSITORIES	160
3.3.1	<i>PRIDE</i>	161
3.3.2	<i>GPMDDB</i>	162
3.3.3	<i>PeptideAtlas</i>	163
3.3.4	<i>Tranche at ProteomeCommons</i>	164
3.3.5	<i>GAPP</i>	165

3.3.6	<i>Human Proteinpedia</i>	165
3.3.7	<i>MAPU</i>	166
3.3.8	<i>PepSeeker</i>	167
3.3.9	<i>SwedCAD and SwedECD</i>	167
3.4	DATA UPLOAD, DOWNLOAD AND FORMAT SUPPORT	168
3.4.1	<i>PRIDE</i>	168
3.4.2	<i>GPMDB</i>	170
3.4.3	<i>PeptideAtlas</i>	170
3.4.4	<i>Tranche</i>	171
3.4.5	<i>GAPP</i>	171
3.4.6	<i>HPP</i>	172
3.5	DATA-MINING AND VISUALISATION	172
3.5.1	<i>PRIDE</i>	172
3.5.2	<i>GPMDB</i>	175
3.5.3	<i>PeptideAtlas</i>	176
3.5.4	<i>Tranche</i>	177
3.5.5	<i>GAPP</i>	178
3.5.6	<i>HPP</i>	181
3.6	DATA CONTENT OF REPOSITORIES IS VARIED	182
3.7	STANDALONE VERSIONS OF SOME PUBLIC REPOSITORIES ARE AVAILABLE	185
3.8	PIPELINES FEED GPMDB, PEPTIDEATLAS AND GAPP DB WITH IDENTIFICATIONS 186	
3.8.1	<i>GPM</i>	186
3.8.2	<i>TPP</i>	188
3.8.3	<i>GAPP</i>	189
3.9	REPOSITORIES FOR QUANTITATIVE PROTEOMICS ARE EMERGING	191
3.10	DISCUSSION: PROTEOMICS DATA QUANTITY, QUALITY AND USAGE	192
3.11	CONCLUSION	195
4	OPTIMISING THE DESIGN OF DECOY SEARCH DATABASES USING THE GENOME ANNOTATING PROTEOMIC PIPELINE (GAPP)	197
4.1	SUMMARY	199
4.2	INTRODUCTION	200
4.3	METHOD	201
4.3.1	<i>Standard datasets with suitable metadata were selected</i>	203

4.3.2	<i>GAPP-APS pipeline produces high quality protein identifications.....</i>	205
4.3.3	<i>Search parameters were applied according to the ABRF metadata</i>	209
4.3.4	<i>Two search strategies were applied.....</i>	210
4.3.5	<i>Nine decoy database designs were investigated.....</i>	211
4.3.6	<i>Decoy database performance was measured using FPR.....</i>	218
4.3.7	<i>Statistical analysis included factorial ANOVA.....</i>	219
4.4	RESULTS	221
4.4.1	<i>FPRs summary.....</i>	221
4.4.2	<i>Recommendation for decoy design based on protein level FPR</i>	224
4.4.3	<i>Recommendation for search strategy based on protein level FPR.....</i>	228
4.4.4	<i>Recommendations based on peptide level FPRs.....</i>	229
4.5	DISCUSSION.....	230
4.5.1	<i>APS threshold explains the differences in FPR between the different decoy database designs</i>	230
4.5.2	<i>Identification performance was comparable with the original ABRF study....</i>	239
4.5.3	<i>Differences in FPR between data submitters was caused by individual sample handling</i>	239
4.6	CONCLUSIONS AND RECOMMENDATIONS TO INCREASE CONFIDENCE IN AUTOMATED SEARCHES.....	247
4.7	ADDITIONAL FUTURE WORK.....	249
4.7.1	<i>Testing further standard datasets.....</i>	249
4.7.2	<i>Apply peptide level reverse decoy database in the public GAPP pipeline</i>	249
4.7.3	<i>Investigation into theoretical FPRs would make the study more useful for non-standard datasets.....</i>	249
5	MRMAID, THE WEB-BASED TOOL FOR DESIGNING MULTIPLE REACTION MONITORING (MRM) TRANSITIONS.....	251
5.1	SUMMARY	253
5.2	BASIC CHARACTERISTICS OF AN SRM ASSAY	254
5.3	THERE IS NO BEST PRACTICE FOR DESIGNING SRM TRANSITIONS	254
5.3.1	<i>Summary of existing transition design software</i>	257
5.4	METHODS	258
5.4.1	<i>An overview of the MRMAid system.....</i>	258
5.4.2	<i>Software implementation.....</i>	260
5.4.3	<i>MRMAid has filters to determine optimal transition candidates.....</i>	262

5.4.4	<i>MRMaid takes a novel approach to transition ranking</i>	269
5.4.5	<i>Transitions can be scored in the absence of MS/MS evidence</i>	274
5.4.6	<i>Results can be downloaded from MRMaid</i>	274
5.4.7	<i>Retention time is calculated using a linear model</i>	275
5.5	RESULTS	276
5.5.1	<i>Man versus MRMaid: testing MRMaid's performance</i>	276
5.5.2	<i>Retention time prediction is accurate</i>	282
5.5.3	<i>MRMaid has comprehensive documentation and user support</i>	283
5.6	DISCUSSION.....	285
5.6.1	<i>MRMaid can design transitions with multiple product ions</i>	285
5.6.2	<i>MRMaid can design multiple transitions for targeting proteins in complex samples</i>	287
5.7	CONCLUSIONS.....	287
5.8	SUGGESTED DEVELOPMENTS FOR MRMAID RELEASES IN THE FUTURE	289
5.8.1	<i>Transition candidate ranking using a star rating</i>	289
5.8.2	<i>Batch mode for submitting protein targets</i>	290
5.8.3	<i>Protein sequence as input</i>	290
5.8.4	<i>Interspecies mode for predicting transitions</i>	291
5.8.5	<i>PSI-compliance of MRMaid</i>	292
5.8.6	<i>Porting data to MRMaid via GAPP</i>	292
6	MRMAID-DB: A REPOSITORY FOR PUBLISHED MRM TRANSITIONS	295
6.1	SUMMARY	297
6.2	INTRODUCTION	298
6.3	AN INTRODUCTION TO MRMAID-DB	299
6.3.1	<i>Existing repositories for SRM transition data</i>	299
6.4	METHOD AND IMPLEMENTATION	301
6.4.1	<i>MRMaid-DB Data Content</i>	301
6.4.2	<i>The MRMaid-DB transitions database schema</i>	302
6.4.3	<i>MRMaid-DB is based on the Biomart framework which offers flexible and federated querying</i>	305
6.4.4	<i>Implementing an instance of Biomart</i>	308
6.4.5	<i>Federating the MRM transitions Biomart with Ensembl Biomart</i>	315
6.4.6	<i>Submitting transition data to MRMaid-DB</i>	316
6.4.7	<i>Exporting data from MRMaid-DB</i>	320

6.4.8	<i>Designing the MR Maid-DB website look and feel.....</i>	320
6.5	RESULTS	321
6.5.1	<i>Examples of MR Maid-DB use cases.....</i>	321
6.6	DISCUSSION.....	326
6.7	CONCLUSIONS.....	330
6.8	FUTURE WORK.....	330
6.8.1	<i>Data content of MR Maid-DB.....</i>	330
6.8.2	<i>MR Maid-DB should become compatible with standards</i>	331
6.8.3	<i>Automated quality control at point of data entry</i>	331
7	CONCLUSION	333
7.1	WIDER OPPORTUNITIES FOR FUTURE WORK	340
7.1.1	<i>Decoy database design options</i>	340
7.1.2	<i>Refactoring of the systems: GAPP and MR Maid.....</i>	341
7.1.3	<i>Automated data harvesting.....</i>	342
7.1.4	<i>Facility to execute different search engines in GAPP.....</i>	343
7.2	FINAL WORD: PROTEOMICS GEAR-HEADS ARE HERE TO STAY	344
	REFERENCES.....	347
	APPENDIX I.....	377
	APPENDIX II.....	383
	APPENDIX III.....	389
	APPENDIX IV	393
	APPENDIX V	405
	APPENDIX VI	409
	APPENDIX VII	413

Table of Figures

Figure 1 The Proteomics Umbrella	7
Figure 2 The number of articles with 'proteomics' in the title or as the topic	8
Figure 3 Experimental flows in mass-spectrometry-based proteomics.....	14
Figure 4 A typical workflow for proteomic mass spectrometry..	21
Figure 5 Schematic diagram of a quadrupole mass analyser	25
Figure 6 Schematic cross section view of an ion trap.....	26
Figure 7 Fourier transform ion cyclotron resonance analyser.....	28
Figure 8 A peptide fragmentation pathway	30
Figure 9 m/z values are applied to peptide identification using search engines	32
Figure 10 A summary of the approaches for quantifying proteins using MS	35
Figure 11 SRM targets a specific peptide to product ion transition.....	38
Figure 12 An overview of proteomic bioinformatics resources for tandem MS.....	42
Figure 13 Examples of MS/MS data in peak list form.....	45
Figure 14 The metadata required by the Genome Annotating Proteomic Pipeline	48
Figure 15 The principle of spectrum identification using PMF and tandem MS searches	51
Figure 16 Peptide assignments from search engines may be filtered by user-specified criteria	59

Figure 17 The problem of protein inference.....	59
Figure 18 Distribution of BBSRC funds in England for bioinformatics research for proteomic MS between 1999 and 2008..	102
Figure 19 Nonlinear Dynamics financial performance	114
Figure 20 Gartner's technology hype cycle applied to proteomics.	126
Figure 21 Investors weigh up biotech start-ups using three main criteria.....	129
Figure 22 Usual funding stages for high-tech start-ups.....	140
Figure 23 Suggested concept of fragmentation of the proteomics value chain.	146
Figure 24 A summary of the benefits and potential uses of public proteomic MS repositories	157
Figure 25 The emergence of public proteomic data repositories has stimulated the development of a huge array of public bioinformatics resources.....	160
Figure 26 PRIDE data visualisation options	174
Figure 27 GPMDB's experiment view	175
Figure 28 PeptideAtlas maps peptides on to the genome using a DAS track.....	177
Figure 29 Homepage for Tranche at ProteomeCommons	178
Figure 30 GAPP's differential view.....	180
Figure 31 Human Proteinpedia's download page.....	181
Figure 32 Increase in data submissions to PRIDE repository	184

Figure 33 An overview of the Genome Annotating Proteomic Pipeline.....	189
Figure 34 Overview of the approach used to investigate the best decoy database	203
Figure 35 Taxonomy of decoy databases	213
Figure 36 Mean protein identification FPR by decoy design and search strategy	222
Figure 37 Box whisker plot to illustrate the distribution of protein level false positive rates (FPRs) across the ten database instances for each decoy design.....	228
Figure 38 Average APS score thresholds across the diverse decoy designs.	233
Figure 39 Understanding differences between decoy designs by examining the number of peptides found by each individual decoy design.....	237
Figure 40 The mean peptide lengths found across ten instances of each decoy type..	238
Figure 41 The distribution of false positive protein identifications across specific Ensembl gene accession numbers, showing the breakdown by decoy type.....	243
Figure 42 The distribution of false positive protein identifications across specific Ensembl gene accession numbers showing the breakdown by data submitter.....	244
Figure 43 False positive rate by decoy design showing the effect of instrument type.....	247
Figure 44 The process of transition design in MRMAid.....	262
Figure 45 MRMAid has three main views for the candidate transitions.....	267
Figure 46 Observed versus predicted retention time for selected transitions	282
Figure 47 Summary of the MRMAid user documentation	284
Figure 48 The process for applying MRMAid to multiple reaction monitoring.....	286

Figure 49 The methodology used to create the MRMAid-DB transitions database.	303
Figure 50 MRM transitions MySQL database schema	304
Figure 51 Dadabik was tested as a possible solution for web-based querying.....	307
Figure 52 The MRM transition database (a MySQL database) shown in Martbuilder.	309
Figure 53 Martbuilder transforms the MySQL transitions database (16 tables) into a large redundant table.....	311
Figure 54 Marteditor software was used to tune the Biomart interface.....	312
Figure 55 Filters and attributes in the MRMAid-DB transitions Biomart.	314
Figure 56 The process of data submission to MRMAid-DB	317
Figure 57 Data submission database tables for the MRMAid-DB website.....	319
Figure 58 Screenshot of results from a federated query using MRMAid-DB.	323

Table of Tables

Table 1 Estimates of the total number of human proteins.....	4
Table 2 Resolution of MS.....	29
Table 3 Standard formats for proteomic MS developed by PSI.....	46
Table 4 A selection of the major protein/peptide identification search engines.....	54
Table 5 Post identification systems.....	62
Table 6 Timeline showing key events in the development of high-throughput proteomics and proteomic bioinformatics.....	83
Table 7 OBT's financial performance.....	107
Table 8 A summary of Matrix Science's products.....	112
Table 9 A summary of Nonlinear Dynamics's products.....	115
Table 10 A summary of GeneBio's products.....	117
Table 11 A summary of Proteome Software's products.....	120
Table 12 A summary of Sage-N Research's products.....	123
Table 13 Funding summary for the proteomic bioinformatics case studies.....	141
Table 14 Comparisons between the company case studies.....	145
Table 15 A summary of data content of the major public repositories.....	183
Table 16 Journals that recommend deposition of MS-based proteomics data.....	185
Table 17 APS_score_info table from the GAPP database schema.....	208

Table 18 The mass tolerance parameters used for each dataset.	210
Table 19 Summary of the database properties of the decoys..	214
Table 20 Mean true positives and false positives across all data submitters and decoy search types.....	223
Table 21 Significant differences between decoy designs for protein identification false positive rate.....	226
Table 22 False positive protein descriptions taken from results of a single instance of each decoy design..	240
Table 23 Three-way ANOVA analysis..	246
Table 24 Overview of empirical methods for SRM transition design applied in the literature to date.	255
Table 25 Optional Boolean filters which may be used to constrain the search for MRM transitions using MRMAid	263
Table 26 Derivation of Transition Score (TS) coefficients..	270
Table 27 MRMAid performance versus experimentally verified transitions taken from Anderson and Hunter, 2006.	277
Table 28 MRMAid performance using horse serum transitions.	280
Table 29 A comparison of the data captured by TraML versus MRMAid-DB	327

Acknowledgements

They say the only place where success comes before work is in the dictionary. Well that's true, but the success of the EngD project is not just down to *my* work. It would have been impossible to deliver relevant research in this interdisciplinary field without constant help and guidance from my colleagues and laboratory-based collaborators.

In particular, I thank Conrad Bessant (pictured), my research supervisor, for giving me the freedom to choose the direction of the research in this EngD. With his guidance, I have achieved more than I thought possible and enjoyed my days - feeling free to try things out. I thank Stephen Regan, my management research supervisor. I am grateful for your insights and editing on my management report; I enjoyed our conversations about poets, bankers and economics.

A huge thank you goes to Luca Bianco, Research Fellow at Cranfield University (also pictured). Luca, a skilful computer scientist, is charming and patient with a self-effacing sense of humour. I have pestered him constantly since he arrived for tips and pointers, and without exception he was always willing to help me and other students in need. He has done a marvellous job with the GAPP pipeline, which needed some serious TLC (and parallel mpi code). Without his help my project would have been a lonely slog and a shadow of what you see here.

I thank Kathryn Lilley (pictured) and Nick Bond at the University of Cambridge for their help with MRMAid and for comments on my EngD reports and research papers. Thank you especially, Kathryn, for the encouraging comments in the bar after the BSPR conference, and for sharing your contacts in the proteomics scene.

I thank my industry supervisors at GlaxoSmithKline, both past (Dan Crowther and Matt Hall) and present (Kieran Todd). Thank you for the support and insights, despite difficult times in the industry.





Thanks also to Chris Barton (pictured on right) and Richard Kay (on the left) at Quotient Bioresearch Ltd, Newmarket. Chris and Rich are my MRM gurus, always willing to answer my questions about lab protocols including MRM, RP-HPLC and MS, as well as acting as guinea pigs testing my prototypes (even on Sundays - thanks Rich!)

I also thank Andrew Spooner for regularly fixing

Apache and MySQL on mo-box (our in-house webserver); Vanessa Ottone for writing the product ion predictor function and implementing the correction factors in the retention time predictor for the MRMAid program; Mike Cauchi for organising lunch time socials, trips to pizza hut and interrupting me for coffee breaks; and Mrs. Kath Tipping (pictured with the EngD students) and Prof. John for coordinating and making EngD Centre funding available for conferences during the MBA/EngD. I also thank my MBA learning team-mates; in particular, I would like to acknowledge the professionalism and great teamwork of Ashish Jain, Mindy Hanzlik and Emmanuelle Clement. Finally, I would like to thank my grandmother, Mary Wright, for her careful proofreading of my thesis. Bioinformatics may not be her preferred mastermind subject, but her meticulous attention to detail and huge vocabulary were invaluable in the polishing of this thesis. I also thank her for regularly beating me at Scrabble.



As you read my thesis I hope you get a sense of how exciting it is to be part of a rapidly expanding research community like bioinformatics. The fact that I have managed to publish five reviews on various aspects of proteomic MS informatics in four years is testament that proteome bioinformatics is truly cutting edge. Of course, I am pretty sure this is the reason; it might be that no one else in the world is willing to sit at a PC for months to write them!

Thank you all! Enjoy and God bless you...

Dedication

*“Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveller, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;”*

From: The Road Not Taken by Robert Frost (1874-1963)

The choice was right. There may have been other options at the time, but these four years on the EngD programme have been stimulating and satisfying, and I am truly grateful for the opportunities I have had. I have felt sheer delight when reviewers or collaborators accept and appreciate my work, and feel ready for the next step in my career.

I would like to dedicate this project to my family: to my own relatives, to those of my new husband, and to my Christian family.

“To whom much is given, much is expected”

Luke 12:48 (New International Version)

I have been blessed with the skills and circumstance to come this far. I pray I will use my experience and career to reflect God’s glory. With God’s help everything is possible...

Abbreviations (science)

ABRF	Association of Biomolecular Resource Facilities
ACN	acetonitrile
API	application programming interface
APS	Average peptide score
AQUA	absolute quantification
CAD	collision activated dissociation
CE	collision energy
CID	collision induced dissociation
Da	Daltons
DAS	distributed annotation system
DB	database
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ESI	electrospray ionisation
FAB	fast atom bombardment
FP	false positive
FPR	false positive rate
FT-ICR	Fourier transform ion cyclotron resonance
GAPP	The Genome Annotating Proteomic Pipeline
GO	gene ontology
GPM(DB)	The Global Proteomics Machine (Database)
GUI	graphical user interface
HPLC	high performance/pressure liquid chromatography
HTP	high-throughput

HPP	Human Proteinpedia
HUPO-PSI	Human Proteome Organisation's Proteomics Standards Initiative
IDA	information-dependent acquisition
IEF	isoelectric focusing
ID-MS	isotopic dilution mass spectrometry
LAMP	LAMP Linux Apache MySQL PHP
LIMS	laboratory information management system
LSD	least significant difference
MALDI	matrix-assisted laser desorption/ionization
MAPU	Max-Planck Unified Proteome Database
MIAPE	minimum information about a proteomics experiment
MIDAS	MRM-initiated detection and sequencing
MRM	multiple reaction monitoring
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MuDPIT	multidimensional proteome identification technique
OLS	ontology look-up service
PHP	PHP Hypertext Preprocessor (scripting language)
pI	isoelectric point
PICR	Protein Identifier Cross-Referencing Tool
PRIDE	PRoteomics IDEntifications database
PTM	post-translational modification
PTP	proteotypic peptide
QQQ	triple quadrupole
QconCAT	Q peptide concatemer

RP(-HPLC)	reverse phase (high performance/pressure liquid chromatography)
RT	retention time
SCX	strong cation exchange (chromatography)
SPMDB	The Standard Protein Mixture Database
SQL	Structured query language
SRM	selected/single reaction monitoring
SSRCalc	Sequence Specific Retention Calculator
TFA	trifluoroacetic acid
TIC	total ion chromatogram
TIQAM	Targeted Identification for Quantitative Analysis by MRM
TP	true positive
TPP	Trans Proteomic Pipeline
TS	Transition Score
TSV	tab-separated values

Amino acid single letter abbreviations

G	Glycine (Gly)
P	Proline (Pro)
A	Alanine (Ala)
V	Valine (Val)
L	Leucine (Leu)
I	Isoleucine (Ile)
M	Methionine (Met)
C	Cysteine (Cys)
F	Phenylalanine (Phe)
Y	Tyrosine (Tyr)
W	Tryptophan (Trp)
H	Histidine (His)
K	Lysine (Lys)
R	Arginine (Arg)
Q	Glutamine (Gln)
N	Asparagine (Asn)
E	Glutamic Acid (Glu)
D	Aspartic Acid (Asp)
S	Serine (Ser)
T	Threonine (Thr)

Abbreviations (business)

CE(E)DD	centre for excellence in (external) drug discovery
CPTAC	Clinical Proteomic Technologies for Cancer
CRO	contract research organisation
GSK	GlaxoSmithKline
OBT	Oxford BioTherapeutics
R&D	research and development
SOP	standard operating procedure
USP	unique selling point
VC	venture capitalist

Thesis Structure Summary

Chapter ^a	Title	Description of content	Publications ^b
1	Introduction	An introduction to mass spectrometry of peptides, bioinformatics analysis in proteomics, and the aims and objectives of this thesis	0
Part I – Characterising proteomic bioinformatics in industry and academia			
2 ^c	The business history of proteomic bioinformatics (1985-2009)	How business and science interacted to create a new industry, proteomics. Investigation into why proteomics and proteome bioinformatics were downsized at the sponsoring company, GlaxoSmithKline.	0
3	Review of public repositories for proteomics	The state of the art in proteomics databases on the internet: including a summary of functionality, data content and data analysis pipelines	3
Part II – Increasing confidence in protein identification using automated analysis			
4 ^d	Optimising the design of decoy search databases using the Genome Annotating Proteomic Pipeline (GAPP)	Investigation into which decoy database design produces the lowest false positive rate for protein and peptide identifications using GAPP and MS/MS datasets of known protein composition.	1
Part III – Software solutions for quantitative proteomics: the MRMAid family			
5 ^d	MRMAid: the web-based tool for designing multiple reaction monitoring (MRM) transitions	Design and implementation of a new tool for automating the design of MRM assays by applying expert knowledge of MRM and MS/MS data-mining	2
6 ^d	MRMAid Database: a repository for published MRM transitions	Design and development of a novel database system for dissemination of published MRM transitions	1
7	Conclusion	Summary of the benefits of this thesis, including contribution to knowledge and a description of possible avenues for future work.	1 ^e
A ^f	References and Appendices		0

a – Each chapter represents a standalone piece of research work, and as such, each has its own brief executive summary. Exceptions to this are the Introduction and Conclusion; **b** – Copies of all published journal papers are available in the Appendices; **c** – Unlike the other chapters, chapter 2 is written for a business/lay audience; **d** - A summary of all the tasks involved in the work presented for these chapters is shown as a breakdown diagram in the corresponding Appendix; **e** – A review of publicly available tools for MRM transition design was written after development of the MRMAid and MRMAid database. This is referred to in the final chapter. **f** - The Appendices are numbered by the Chapter to which they refer. In Appendix I there is a summary of the courses/conferences attended by the author as part of the EngD programme, and membership to professional organisations.

Introduction

1.1 Proteomics is the large-scale study of proteins

Proteomics is the science concerned with understanding the role of protein molecules in biology, such as their function in maintaining health, causing disease and in development and ageing. Proteomics techniques aim to measure and characterise proteins present in cells, tissues or whole organisms under a set of defined conditions, at a particular point in time.

Protein molecules underpin living processes in every cell in every organism on Earth. They provide cellular machinery to maintain life providing structural supports, acting as signal messengers/transducers and reaction catalysis. Proteins are polymers, since they consist of many individual amino acids molecules covalently bonded together into long chains. The chemical/physical properties of the amino acids in the chain, their position, and the cellular environment of the protein act together to determine how the final protein will fold up into its final 3D structure. The resulting structure dictates the functional role of the protein, how it interacts with other biomolecules, where it is located, and the substrates it can bind.

Proteomics involves characterising the ‘proteome’: the complete set of proteins expressed in a specific tissue, organ or cell type. Unlike the genome, a DNA molecule that is a static string of nucleotides from birth to death, the proteome is highly dynamic, since proteins carry out virtually all cellular functions and respond to constantly changing intra- and extracellular environments.

Indeed, it is estimated that there are 20,488 genes that encode proteins in the human body (Table 1), but the real size of the proteome depends on the measure taken; for example, if all splice variants are taken into account, the number of distinct proteins is estimated to be over 200,000 (Table 1).

Table 1 Estimates of the total number of human proteins. Data taken from (Uhlen and Ponten, 2005, Clamp *et al.*, 2007)

Description	Number of proteins
If one protein is counted for each gene locus	20,488
If protein fragments (such as splice variants) are counted	>200,000
If proteins that differ in post-translational modifications are counted	<100,000
If proteins that differ by small genetic variations (such as single nucleotide polymorphisms)	>75,000

This huge variation in the composition is dependent on the individual’s state of health and age, gender, as well as on genetic differences, such as race, mutations, and other factors. Consequently, proteomics research (at least for now) must focus on characterising proteomes of very specific samples under well-defined conditions.

The word 'proteomics' is an 'umbrella' term that refers to a diverse array of experimental approaches; each employing its own set of niche technologies (Figure 1). For example, proteomics research can involve structural studies, interactomics and identification/ quantification of proteins using mass spectrometry (MS).

The benefit of proteomics is that it is a direct method. Compared to indirect measurements, such as gene expression analysis via DNA microarrays, for example, proteomics can reveal real changes in proteins themselves. It is proteins that carry out the vast majority of functional and structural roles in cells, not the genetic information that underlies them.

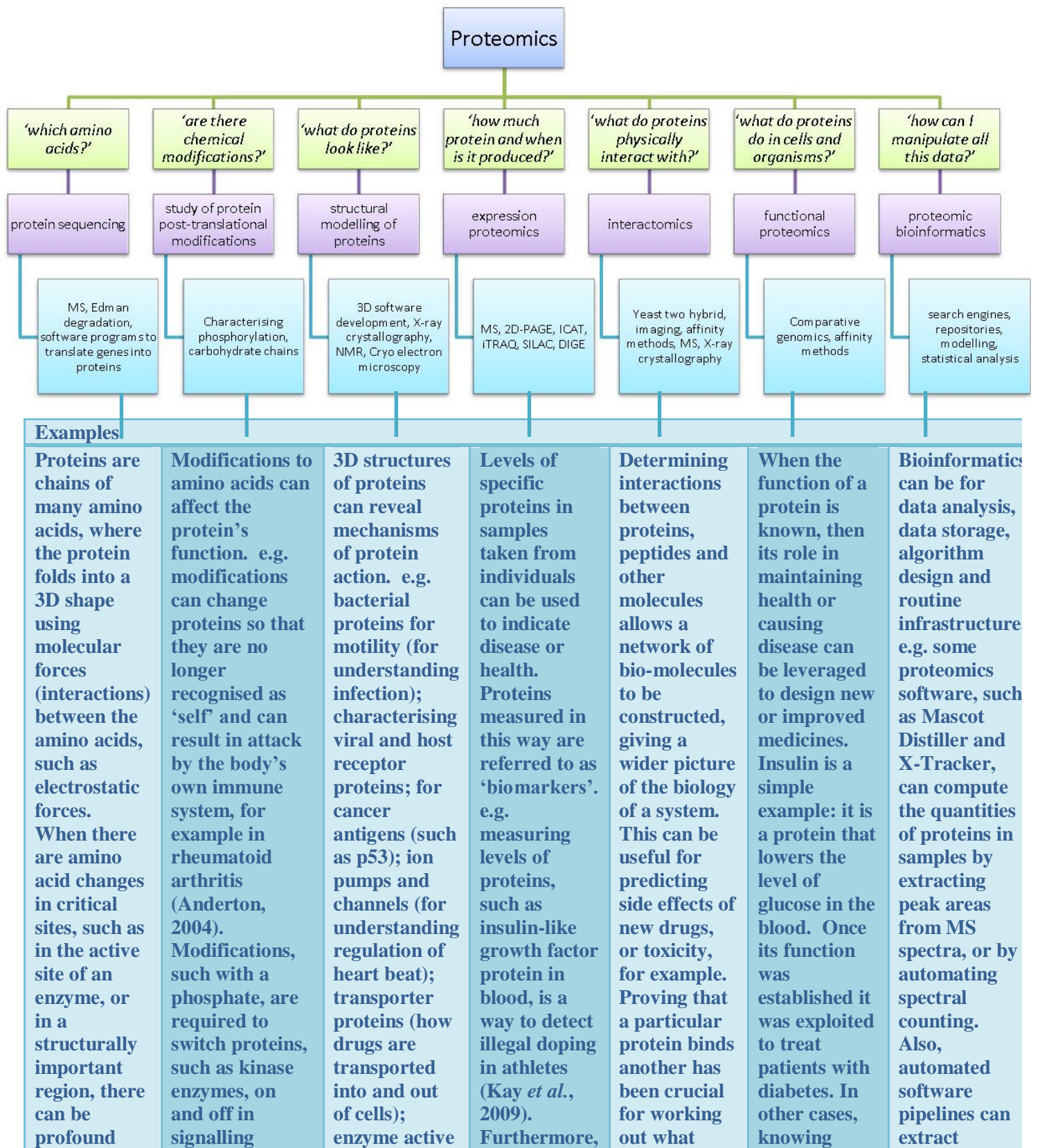


Figure 1 The Proteomics Umbrella. ‘Proteomics’ describes any large-scale approach taken to investigate the function and properties of proteins (biology’s molecular machines). The ‘proteome’ is the entire complement of proteins expressed at any one time, in a particular system, cell type or tissue under defined conditions, and it can be examined from very different angles to answer very different questions. To reflect the diversity of research activities undertaken, proteomics can be split into separate branches of research with its own specific toolbox of experimental approaches and methods. The broad categories shown are not mutually exclusive and there are activities which do not fit in those areas shown here. One of the main aims of proteomics is to discover which proteins indicate a disease state or susceptibility to a disease, so-called protein biomarkers. These can be used for designing new drugs and in some cases can prove useful for demonstrating to the regulators that a drug works. (Source: authors own summary)

1.1.1 Proteomic mass spectrometry is an approach used for biomarker discovery

The current arsenal of drugs that the pharmaceutical industry offer target only several hundred or so distinct proteins. Through proteomics approaches, however, the potential to expand on this number was widened dramatically. Indeed, as a field, proteomics has grown rapidly in the last decade (see Figure 2); in the late 1990s to 2000s, for example, academic researchers and biotech/pharmaceutical companies became particularly interested in a high-throughput approach for identifying proteins in biological samples, namely MS-based proteomics. This was because it offered a new avenue to discover biomarkers that have potentially high commercial value.

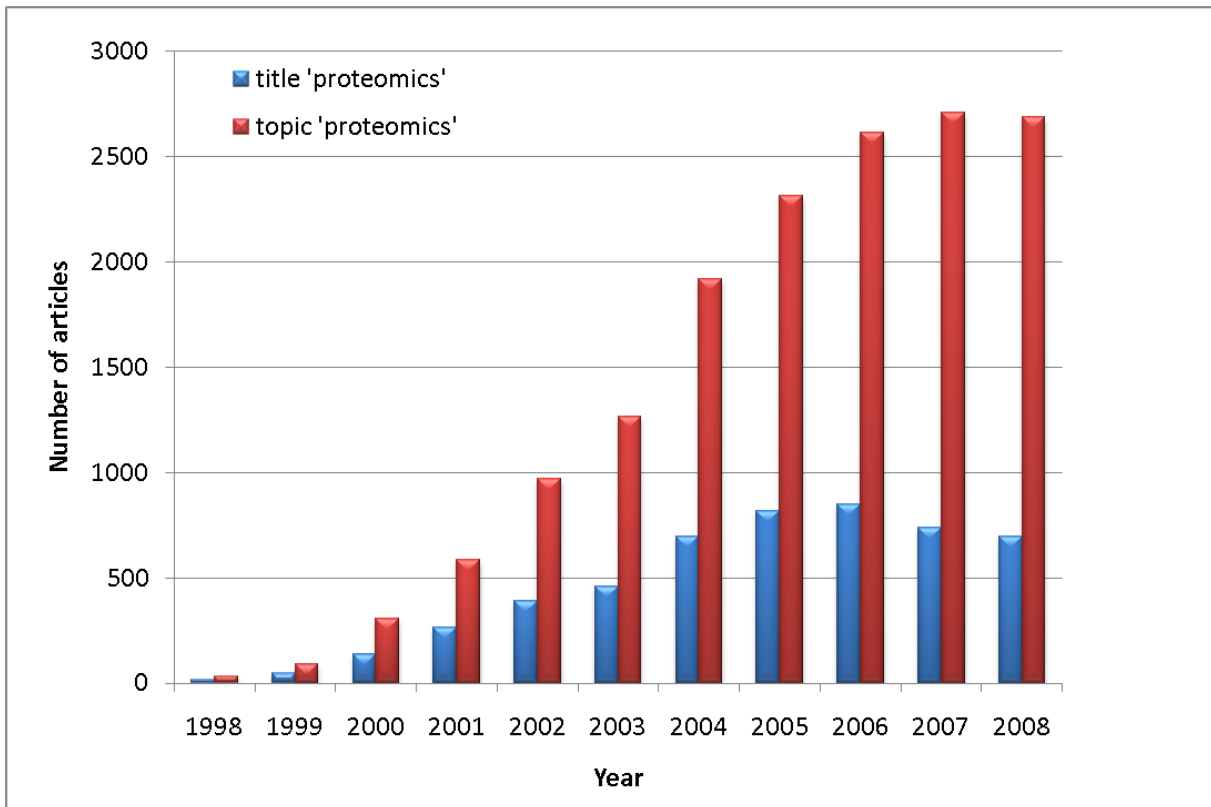


Figure 2 The number of articles with ‘proteomics’ in the title or as the topic (source: ISI Web of Knowledge, Science Citation Index Expanded (SCI-EXPANDED) database)

By comparing protein profiles identified by MS for samples from diseased versus healthy populations, for example, biomarkers for elucidating disease mechanisms and for developing diagnostics and new therapies can be identified. These new targets would then be patented as new drug targets. Conditions including cancer, atherosclerosis, pain, virus-induced cell transformation and many others have been investigated using this approach, although biomarker discovery turned out to be a complex task, as is explained further in Chapter 2.

1.1.2 Proteomics is an ‘omics’ science that is needed for systems biology

Another benefit of large-scale protein studies has emerged recently, that is proteomics as a branch of ‘systems biology’. A major aim of biology research is to achieve a complete computational model of the cell: a ‘cell simulator’, which can be used to demonstrate the effects of drugs, environment and genetics on biology to understand disease, human development and more.

Modern drug design in the pharmaceutical industry is not as ‘rational’ as scientists would like. One cannot, for example, predict how a drug will affect the metabolic processes in a cancer cell until it is empirically tested using a cell line or animal model, because there are no holistic models to simulate it. Thus, only by characterising the changes in molecules in time and space, using so-called ‘global’ approaches, such as genomics (study and prediction of genes), transcriptomics (transcripts i.e. mRNAs which encode proteins), proteomics (protein expression), interactomics (molecular partners, complexes, dynamics) and metabolomics (small molecules and metabolites), can there be any hope of generating a mathematical model of a cell. These large-scale ‘omics’ approaches are required, since so many discrete events and interactions are happening simultaneously. If systems biology ever arrives in this global form, where computational models are available for whole systems or even whole organisms, then the way medical research is performed and

demonstrated to the regulators will be totally revolutionised; changed beyond all recognition.

Great advances have been made to get this far: for example in the 1970s, it could take an entire PhD project to sequence a single gene using lab-based techniques, but by mid-2000s it was possible to sequence the genome of a whole organism in a day (PressRelease, 2007i). Compared to the biology being studied before, this scale of research is revolutionary. The leap in throughput became possible, in this case, through the advent of: new methods in molecular biology (invention of polymerase chain reaction (PCR), cloning and sequencing methods); advances in instrumentation (automated sequencers, robotics); and computing and software development (sequence recombination algorithms, processing power). The trends seen in the development of proteomics also relied on significant technological breakthroughs, as explained later.

Unsurprisingly, there is much work to do before researchers can unravel what this new 'omics' data really means in the context of an organism, tissue and at the level of the cell (Ghosh and Poisson, 2009). Nevertheless, the desire to achieve the elusive systems model for biology acts as a potent stimulus for proteomics and proteome-related computational research, including for the research presented in this thesis.

1.1.3 Proteomic bioinformatics exploits computers for analysing large datasets from proteomic MS experiments

The promise of proteomics - to deliver biomarkers and new ways to model processes in biology, for example - is wholly dependent on computers, and their ability to manipulate large datasets in an automated fashion.

When computers are applied to analyse biological data of any kind, the process is referred to as 'bioinformatics'. Bioinformatics is an interdisciplinary science that sits at the interface between the biological and computational sciences. This thesis delivers bioinformatics solutions specifically for proteome research, so the work is described as 'proteomic bioinformatics', or 'proteome bioinformatics'. In particular, the work presented here is focused on delivering computational methods and tools for *mass spectrometry*-based proteomics that may be applied directly to early stage pharmaceutical and medical research.

To put the project deliverables into context the next sections introduce some of the major concepts in proteomic MS and proteome bioinformatics. The story of how proteome bioinformatics approaches and resources came about is presented in Chapter 2, and a detailed review of the state of the art in public computational resources for proteomics is presented in Chapter 3; however, the following section gives the reader a brief introduction to the field, so the contribution to knowledge and benefits of this thesis can be appreciated.

1.2 The practicalities of high-throughput proteomic MS and proteome bioinformatics

1.2.1 Proteins are prepared for MS using 2D-PAGE or MuDPIT

MS requires that peptides enter the MS in a charged and gaseous state, preferably only one peptide species at a time. There are two main routes to achieving this (Figure 3): (1) two dimensional poly-acrylamide gel electrophoresis (2D-PAGE), and (2) the multidimensional proteome identification technique (MuDPIT). The first approach is still practised, but is less routinely used for high-throughput studies; both are explained now.

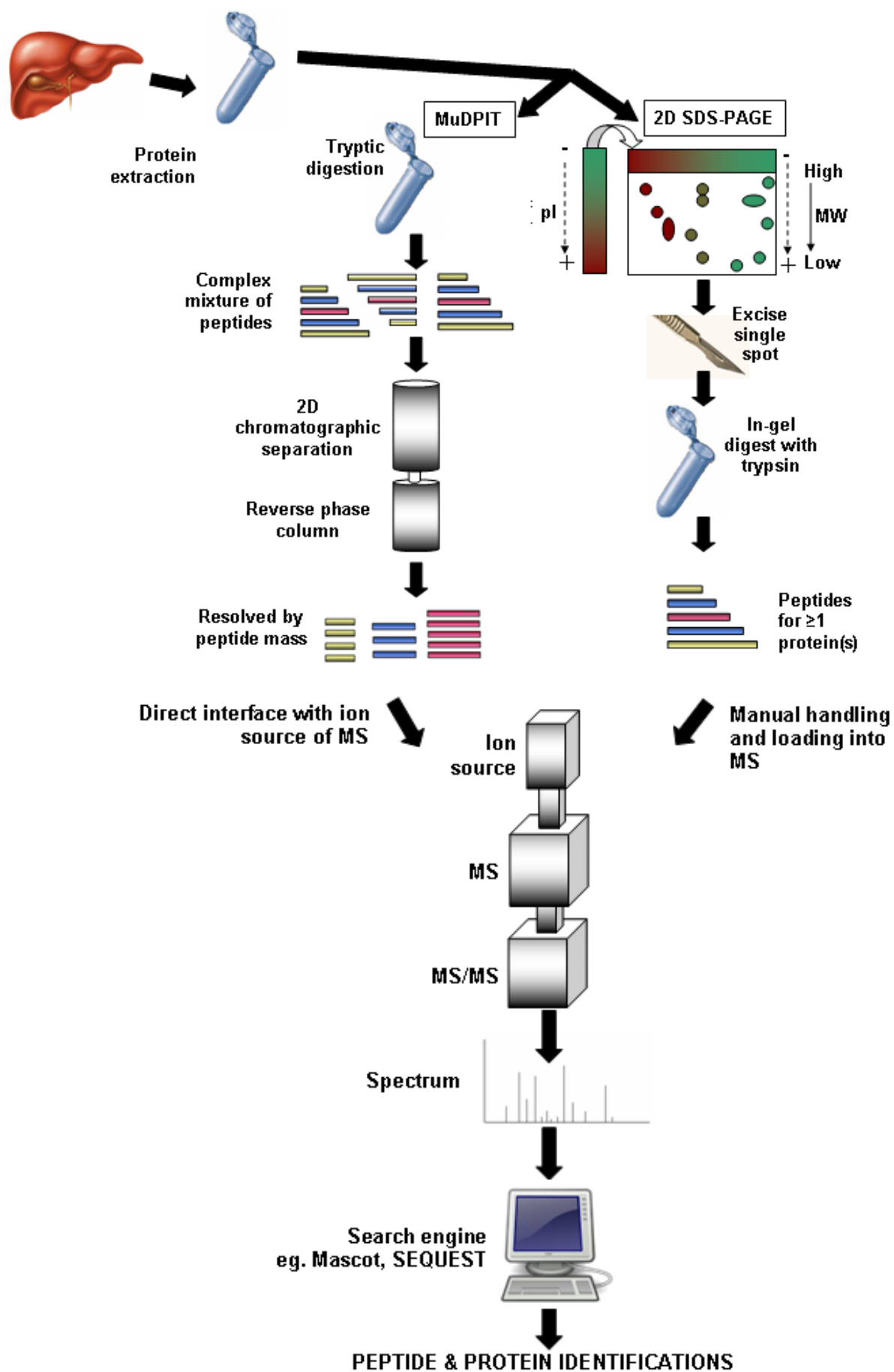


Figure 3 Experimental flows in mass-spectrometry-based proteomics. MuDPIT (multi-dimensional protein identification technology) uses two chromatography steps interfaced back to back, the advantage of this being band broadening and increased resolution. The capillary can also be fed directly into the ion source of the MS to maximise sensitivity. 2D-SDS PAGE separates proteins by two different properties: pI and molecular weight. Picked spots may also undergo an HPLC stage after digestion to increase purity and to allow automated introduction into the MS instrument.

Gel-based methods separate proteins by their chemical and physical properties

2D-PAGE is a gel electrophoresis approach for separating complex protein mixtures. It separates protein molecules by two different properties: mass and isoelectric point (pI). Protein molecules can have a positive, neutral or negative charge, depending on the chemical groups in the molecule and the chemical environment; pI is the pH value at which the net charge across the whole protein molecule is zero.

To create the gel, the proteins are first separated by pI in an isoelectric focusing (IEF) step. The proteins are put on a strip with a pH gradient and a voltage difference is applied. The proteins are free to migrate according to their charge state, moving towards the pole with an opposite charge until the protein reaches the location where its pI equals the strip's local pH, where it halts because its net charge is zero. The proteins are now separated by pI. Next, they are separated by mass by rotating the strip by 90 degrees and separating using sodium dodecyl sulphate (SDS) PAGE.

SDS denatures proteins - straightening them all out into linear structures - and imparts a uniform negative charge, with the number of charges adopted being dependent on the protein's length. To explain: when unfolded, a protein's length is approximately proportional to its mass, thus each protein attaches a number of SDS molecules approximately proportional its mass. Since the SDS molecules are negatively charged, the result is all proteins having approximately the same mass-to-charge ratio as each other. As mentioned, SDS also imparts a negative charge which is necessary here, because the proteins must become charged (having lost their charge in the IEF step) to migrate across a potential difference. The unfolded proteins have to pass through the gel matrix, whereby larger proteins travel slower than smaller ones.

Now separated by pI and mass, the proteins are stained whilst in the gel with a dye, the result being a canvas of spots, where each protein species has a unique position on the gel 'map': its x,y-coordinates depending on its unique properties. The protein spots are usually excised and then proteolytically digested into peptides. The proteolytic enzyme of choice is trypsin, which cleaves the carboxy (C)-terminal of arginine or lysine generating peptides with at least one basic amino acid, thus trypsin predominantly generates positively charged peptide ions - a property which

is very useful, since only charged ions can be analysed in MS, and positive ions are the ions of choice. The digested protein samples are then ready for ionisation in MS.

MuDPIT exploits high performance liquid chromatography to separate peptides

2D-PAGE has a protein or peptide-picking stage, where peptides are individually excised from the gel and dissolved in buffer before analysis in MS; this process is expensive to automate, using robots for example. In contrast, MuDPIT (Washburn *et al.*, 2001) offers the advantage that it may be fully automated at relatively low cost. Compared to 2D-PAGE, when using MuDPIT, the IEF step (first dimension) is substituted by strong cation exchange (SCX) chromatography, and the separation by mass (second dimension) is replaced reverse-phase (RP) HPLC. For MuDPIT, the protein mixture is digested into peptides first and then the pooled peptides are separated and purified using capillary chromatography.

In chromatography, the mixture of molecules to be separated is dissolved in a 'mobile phase' which is passed through a 'stationary phase', usually immobilised on a column or on beads packed into a column. Certain molecules become separated from other molecules in the mixture based on differential partitioning between the mobile and stationary phases. And the mobile phase moves through the column with the help of an HPLC pump.

SCX chromatography

Ion-exchange chromatography has a charged stationary phase, thus separation of peptide ions is achieved by differing levels of electrostatic attraction between the ions and the column. Ions of the same charge state as the column and uncharged compounds are ejected, whereas ions with opposing charge states attract and bind, then must be eluted off gradually. The SCX stationary phase is usually sulphonic acid polymer, as explained in (Zhang *et al.*, 2003) for example, immobilised on silica particles (usually 5µm diameter). Peptides are retained by the surface charge of the stationary phase because, as mentioned, most peptides derived from digestion with trypsin contain the basic amino acids lysine and/or arginine, so at the pH applied for SCX (often pH2) most peptides are cationic, with two or three positive charges. Furthermore, organic solvent (such as acetonitrile (ACN)) is used to strengthen the ionic and hydrophilic interactions, facilitating stronger retention of these charged peptides.

To gradually elute off the bound peptides, an increase in counter ion concentration is needed; the pH may be increased to do this, or the salt content of the buffer may be increased at constant pH (e.g.pH3) starting at low concentration (e.g 10 mM) then

increasing (e.g. to 1M) ammonium formate, for instance. In most cases, the mobile phase is usually an organic solvent, such as 5% formic acid in 20-50% ACN.

RP-HPLC

After SCX, peptides enter RP-HPLC. RP-HPLC has a hydrophobic stationary phase and an aqueous polar mobile phase. For peptide separation, the stationary phase is usually silica beads coated in a hydrophobic straight-chain alkyl group, $C_{18}H_{37}$ - referred to a 'C₁₈'. Peptides bind to the beads in high aqueous mobile phase, and elute off with high organic mobile phase (such as ACN); the peptides are usually all eluted by 50% organic solvent. Since the binding of peptides to the beads is based on hydrophobic forces, thus retention time (RT) increases with hydrophobic (non-polar) surface area of the peptides.

In practice, peptides are separated by running an ascending gradient of an organic solvent over time that is then brought back to the starting concentration over about a specified time period. The gradient may be performed in a linear or in stepwise fashion.

Examples of solvents that are used are: A (aqueous, water with 0.1% acid) and B (organic solvent with 0.1% acid), whereby the acid (such as trifluoroacetic acid (TFA) or formic acid) is added to improve chromatographic peak shape. The sample

containing the peptides is usually dissolved into phase A. One example of a gradient is starting the mobile phase with A at 98% and B 2%, then increasing B to 60% of the total mobile phase composition and bringing it back down to 2% again over 60 minutes. In this case, most peptides will elute at around 30% organic solvent, and if there are very hydrophobic ones these will elute nearer 60%. The gradients applied in RP-HPLC are extremely variable between researchers, and often require optimising, but the overall aim is the same: to have distinct peptide species elute at discrete time points into MS.

In RP-HPLC, there are many factors that can influence RT of peptides, including: particle size, column dimensions (length and diameter), flow rate, temperature, pH and mobile phase composition and gradient. For example, C₁₈ beads with a diameter of less than 2 µm can offer significant increases in throughput for peptide analysis (Kay *et al.*, 2007), although beads of 2-3 µm are more routinely used in proteomics. Flow rates may vary, including nanoflow (nl/min), microflow (µl/min) and normal flow (ml/min), depending on the column used; the advantage of low flow rates being that greater sensitivity can be achieved (Kay *et al.*, 2007). Temperature is usually maintained at 30-40 °C throughout separation, depending on the experiment.

RP-HPLC is used frequently, and there are software tools for predicting theoretical RT of peptides in HPLC for researchers planning proteomics studies. The

algorithms are based on the assumption that the chromatographic behaviour of peptides is dependent predominantly on their amino acid composition, so by summing the hydrophobic contribution of each residue, the RT can be predicted. A notable example, Sequence Specific Retention Calculator (SSRCalc) (Krokhin *et al.*, 2004), applies a linear model that can accurately predict RT for peptides up to approximately 20 residues, and correction factors are specified in the prediction process to account for the variability in column set-up. In addition, neural networks may be applied to RT datasets for the purpose of optimising models for RT prediction (Petritis *et al.*, 2003).

1.2.2 Mass spectrometry of peptides can be performed in high-throughput

MS was originally a method used in small molecule chemistry, since it allowed elucidation of chemical structures by measuring a molecule's m/z : mass (m) divided by its charge state (z). From the m/z values one could determine the atoms involved and hence identify basic structures. Over the decades, MS has evolved and can now routinely achieve resolutions in excess of 0.1 Daltons² with the latest instruments and has become more routinely applied to analyse large biomolecules, such as proteins (see Figure 4 for a typical workflow).

² A Dalton is the same as an atomic mass unit; it is the usual way to express mass when referring to proteins

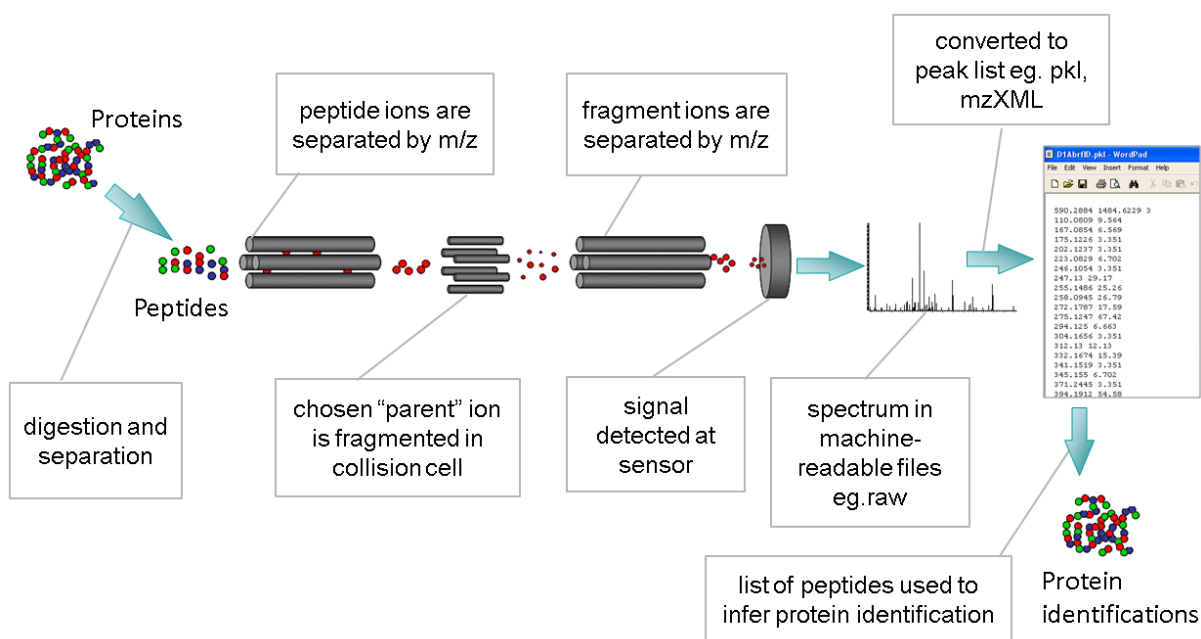


Figure 4 A typical workflow for proteomic mass spectrometry. Notice that data straight from the MS instrument is in 'raw', machine-readable form, meaning it usually requires MS-vendor-specific software to convert it to a peak list of intensities and m/z values. This conversion step often requires proprietary software or software libraries.

There are two common modes of mass spectrometry: single MS or tandem MS (also referred to as MS/MS, or MS²). The former produces a spectrum of the peptide ion m/z values, the latter produces a spectrum of peptide *fragment* ion m/z values. Tandem MS is the approach used most frequently for proteomic MS. Various mass spectrometers are available, each setup offering different benefits and applications; some of these setups are explained in detail later in the thesis, such as triple quadrupole (QQQ) MS, and others. The general principles of MS/MS are now described.

A tandem mass spectrometer has three components: an ion source, a mass-to-charge ratio mass analyser, collision cell (a type of analyser), and a detector.

Ion source

The ion source is at the entrance of the instrument and its function is to ensure the molecular species entering the analyser are charged and in the gas phase. The most common ion source in proteomic MS is electrospray ionisation (ESI) (Fenn *et al.*, 1989), followed by matrix-assisted laser desorption/ionization (MALDI) (Tanaka *et al.*, 1988).

Using MuDPIT, the separation phase may be directly coupled to the MS instrument, whereby the continuous flow of peptide-separated liquid sample is injected through a fine nozzle into the instrument. Once inside, the peptides in the liquid undergo ESI to atomise the peptides. ESI is performed at the tip of a very fine nozzle (needle), which is heated (to around 60°C), at atmospheric pressure, and a drying inert gas is applied over the tip. The nozzle is conductive: a potential difference of up to 5kV is applied. This means the positively charged peptides are pushed out of the needle tip, and as the liquid containing the peptide – which is more volatile than the peptides – dries off, the droplet containing the peptide ions gets smaller, until at a critical moment when the charge density destabilises the droplet and the peptide ions repel each other dispersing as a fine aerosol. The vaporised ions then drift towards the opposing needle electrode.

ESI does not impart a charge on to the peptide; it is a 'gentle' technique, whereby the native charge state of the molecular species is exploited. Most peptides will have a positive charge if derived from tryptic proteolysis, so the technique is particularly suited to proteomics. Native states of molecules are often preferred for biologists wishing to characterise nature.

Micro and nano-spray are recent variants of ESI that are employed in proteomic MS studies. The 'micro' and 'nano' prefix refer to the flow rates of liquid in RP-HPLC into the instrument; by microlitre (μl) or nanolitre (nl) per minute, respectively.

MALDI is the other ionisation technique employed for proteomic MS. It employs a laser, usually UV light, which is directed at a metal plate onto which a 'matrix' with spots of peptide (or whole proteins) have been immobilised. The matrix consists of crystallised molecules³ such that when the energy from the laser hits the matrix, the matrix itself becomes ionised. Part of its acquired charge then transfers to the analyte molecules (the peptides). In this way, the peptides become ionised (usually by addition of a proton, thus $[\text{M}+\text{H}]^+$), but are protected from the disruptive energy of the laser. MALDI usually produces singly charged ions and, in contrast to ESI, is usually performed in a vacuum.

³ The most common matrix constituents are 3,5-dimethoxy-4-hydroxycinnamic acid, α -cyano-4-hydroxycinnamic acid and 2,5-dihydroxybenzoic acid.

Analyser

The analyser measures the mass to charge (m/z) ratio of ions⁴. The two main strategies to achieve this are by using an electrical field (such as in time-of-flight (TOF), quadrupole and ion trap analysers), or using a magnetic field (in the Fourier transform ion cyclotron resonance (FT-ICR) analyser).

Data from various instrumental setups has been applied to develop the novel tools and methods in this thesis; in particular, the **quadrupole** analyser data, which is the analyser of choice for selected reaction monitoring (SRM) (Chapter 5 and 6). Also, MS instruments with **ion trap** and **Fourier transform ICR** analysers have been used to produce the standard datasets applied in Chapter 4, and to populate the database (GAPP DB) interrogated by the novel tool presented in Chapter 5. The principles of these three analysers are described now.

The quadrupole (Paul and Steinwedel, 1953) has four conductive, metal rods aligned in parallel arranged in a square (Figure 5). An oscillating electrical field is produced in the chamber by the application of a direct-current and a radio-frequency alternating current. When the peptide ions enter the field, they oscillate perpendicularly to the direction of their movement, with their progress in the field determined by the ions' charge and mass. Only for certain mass and charge

⁴ m/z means mass divided by charge, for example a peptide of 670 Da and a 2+ charge would have an m/z of 335

combinations will an ion successfully escape from the quadrupole, all others will have increasing horizontal or vertical amplitudes and will be ejected before the end. The quadrupole can be tuned scan for a narrow or a wide range of m/z values.

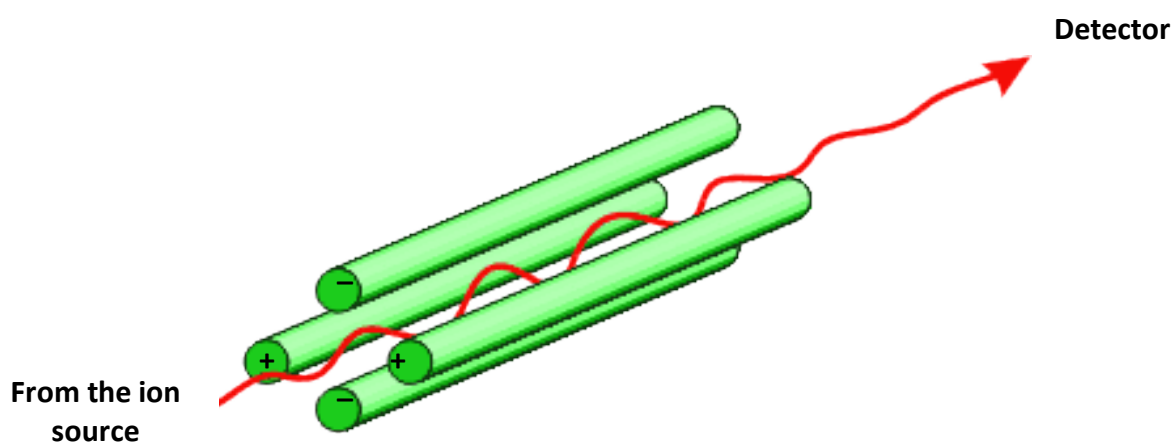


Figure 5 Schematic diagram of a quadrupole mass analyser. The red arrow shows the path of selected ions (Source: adapted from Wikimedia Commons)

Ion traps (March, 1996) are a type of ion cage that store charged particles by the use of an oscillating electric field (in a similar fashion to a quadrupole). An ion species is stored in a vacuum chamber between caps of the same polarity; there are three electrodes: two capping electrodes (one connected to the ion inlet and the other to the detector plate), and a single ring electrode that encircles the trap chamber (Figure 6).

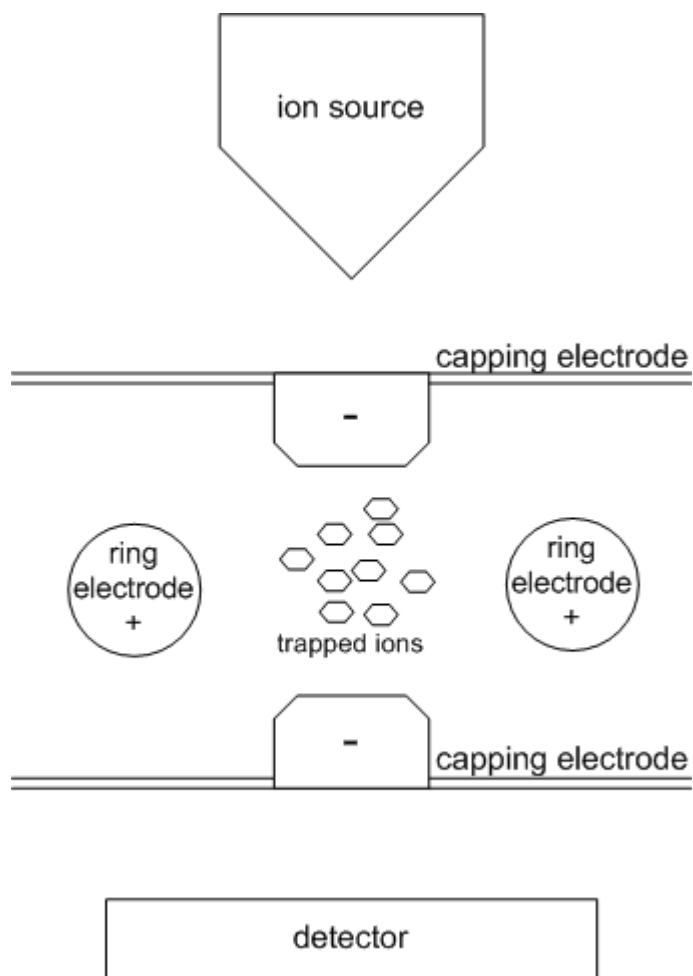


Figure 6 Schematic cross section view of an ion trap (Source: adapted to a schematic from the Paul ion trap entry at Wikimedia Commons)

The ions enter the trap, where there is an applied oscillating radiofrequency electrical field (like in a quadrupole). This field causes the ions to shift and move according to the applied field, producing a compact ‘cloud’ of ions that expands and compresses. To dissipate the energy generated from ion collisions, an inert gas, such as helium, is continuously added to the chamber. The m/z can be measured by an ion trap because the ring electrode potential difference sets the m/z threshold below which ions are expelled from the trap. As for the quadrupole, the voltage setting may be configured to accept a range of m/z values.

The 'cyclotron frequency' of ions is the measure used to determine m/z of peptide ions in FT-ICR (Figure 7); space and time measures, as for other analysers, are not used. It is the highest resolution, and highest cost approach to proteomic MS because it requires a super-conducting magnet.

The analyser is a box-type compartment with a plate on each side: encircling the chamber are two 'excitation' plates (opposite to each other) two 'detector' plates (also opposites), and the entrance and exit plates are the 'trapping' plates. When ions enter the magnetic field in this box, they are forced into a circular trajectory by the plates, thus becoming trapped inside. An alternating current (as radio frequency) is applied to the excitation plates, which causes ions trapped in the magnetic field to become excited to higher energies, and trajectories. At the moment when the current is stopped the ions decay back to their original states. The trajectories of the ions in the field induces an electric current that is measured by the detector plates. It is this current that is measured. It corresponds to the decay in kinetic energy of all ions in the trap, and may be deconvoluted into mass-to-charge ratios by applying the mathematical operation, Fourier transformation.



Figure 7 Fourier transform ion cyclotron resonance analysers use high-maintenance superconducting magnets (Source: Wikimedia Commons)

The mass accuracy and general performance achieved by each type of instrument varies enormously (Table 2).

Table 2 Resolution of MS (Source: adapted from (Domon and Aebersold, 2006))

Characteristic of the MS setup	IT-LIT	QQQ	QQ-LIT	FT-ICR
Mass accuracy	low	medium	medium	very high
Resolving power	low	low	Low	very high
Sensitivity	high	high	High	medium
Dynamic range	low	high	High	medium
Throughput	high	medium	medium	medium
Suitability for peptide identification	medium	low	Low	high
Suitability for absolute quantitation	low	high	High	medium
Suitability for detection of PTMs	low	low	High	low

Collision cell

In tandem MS mode, the peptide ions are broken into smaller peptide fragment ions in a collision cell (or 'chamber'), which is often an ion trap or quadrupole, containing inert gas, such as helium or argon. An electrical potential is applied to the cell, imparting kinetic energy to the particles. This kinetic energy is converted into internal energy in the peptide molecular ions resulting in bond breakage and the release of smaller fragment molecules. The collisions are stochastic, so even for replicate samples different ions may be observed.

The primary mechanism of peptide fragmentation, regardless of the ion source used, is collision induced dissociation (CID) - also referred to as collision activated dissociation (CAD)⁵. In CID, there are six primary ion types created during peptide fragmentation pathways; these are referred to as b-, y-, a-, x-, c- and z-type ions. The

⁵ Post source decay is another mechanism for peptide decay, where peptide ions spontaneously decay into fragments in the ion source vacuum. It is a phenomenon specific to the MALDI ion source.

formation of b- or y-ions, the most common fragment ions, is shown in Figure 8. The original paper describing the nomenclature and pathways is (Roepstorff and Fohlmann, 1984).

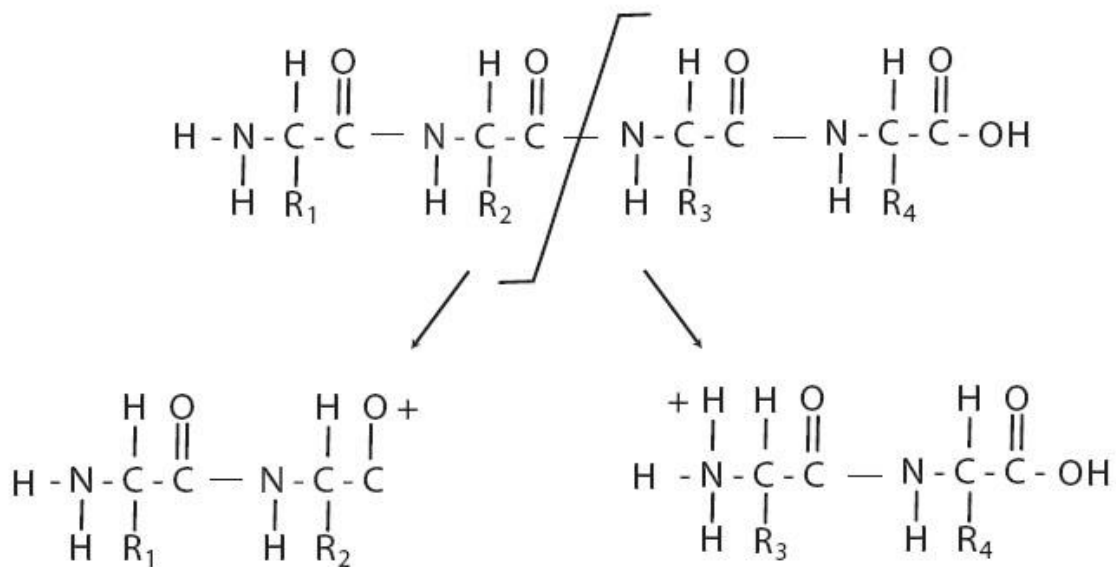


Figure 8 A peptide fragmentation pathway. Frequently the peptide bond breaks to generate fragments, only one of which retains a positive charge and is detected in MS: either a b-ion (left, retaining the amino terminus) or a y-ion (right, maintaining the carboxyl group) (Source: (Brunetti *et al.*, 2008))

Molecular ions that retain a positive charge are fed into a final mass analyser, where they are focused using a magnetic field and their signal intensity measured at the sensor.

The propensity for specific fragmentation pathways depends on the MS instrument in use. Triple quadrupole MS, for example, has a bias towards detection of y-ions,

rather than b-ions. Moreover, the complement of ions produced is also dependent on the voltage applied to the cell and pressure of the inert gas, properties which can be altered by the operator. For example, 'high-energy' CID generates spectra that usually from single collisions between peptide ions and the gas particles, but generally results in a wider range of fragmentation possibilities. In contrast, 'low-energy' mode permits several collisions and hence more complex, multi-step cleavages are observed, but the overall results is a greater proportion of b- and y-ions, and fewer of the other species of ions seen in high-energy CID.

Detector/sensor

The detector measures either the current produced or the charge induced when an ion passes or directly hits a surface. Since the volume of ions exiting the mass analyser at any given moment is small the signal must be amplified. Microchannel plate detectors are frequently fitted in modern instruments, these are a type of electron multiplier.

In FT-ICR and Orbitraps (modified ion traps), ions are not detected by hitting a detector, such as an electron multiplier, but rather are measured by passing near detection plates. No current is produced, only a weak AC image current is detected in a circuit between the electrodes.

The output at the detector is a mass spectrum, usually representing the complement of fragment ions for a single peptide precursor ion, with m/z on the x-axis and arbitrary signal intensity on the y-axis (Figure 9).

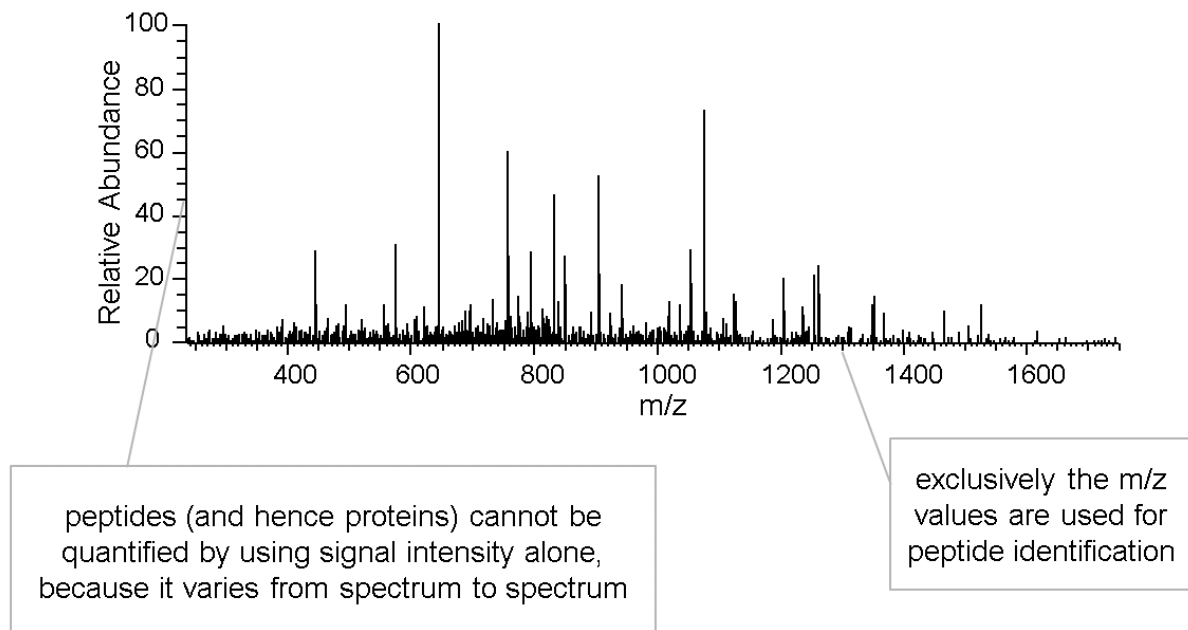


Figure 9 An MS spectrum. Only the x-axis m/z values are applied to peptide identification using search engines

Each spike represents a single fragment ion. The identity of the peptides may be determined manually by inspection of these product ion spectra, however, this is tedious and in many cases the volume of data precludes the manual approach. Therefore, an automated proteomic search engine is usually applied to produce protein identifications from MS/MS peak lists, see section 1.2.4 for details of this process.

Visibility of peptides in MS

Not all peptides can be detected using MS with the aforementioned approach. There are four aspects that determine whether a peptide will be observed or not in MS (Tang *et al.*, 2006):

- (i) Chemical properties of the peptide – will the structure successfully undergo ionisation at the ion source? Will it fragment in the collision cell to produce detectable product ions?
- (ii) Limitations of the peptide identification protocol, including the pre-processing of the sample - Is the peptide positively charged? Is tryptic digestion complete so peptides are within mass range?
- (iii) Abundance of the peptide in the sample – is there enough peptide to be detected? (Of course, there is no PCR for proteins to amplify the sample)
- (iv) Interference with other peptides present in the sample – is there competition with another peptide in the identification procedure? Is there co-elution of peptides making signals difficult to interpret?

Some of these factors are very hard to predict, but given the data the community already has, it is clear that certain peptides are routinely more readily detectable than others, given standard MS protocols. Peptides that are usually the ones that are visible for a given protein in MS are referred to as proteotypic peptides (PTPs) for the given protein.

Given that peptides are usually measured in MS with a view to identifying the parent protein, in some definitions 'proteotypic' refers to those peptides that are both visible *and* unique for a given protein. In this EngD thesis, both constraints on the definition are applied: usually visible and unique to the protein.

There are various computational/mathematical approaches for predicting PTPs for the purposes of proteomic MS. These include: neural networks (Tang *et al.*, 2006); classical pattern discovery methods (Mallick *et al.*, 2007); machine learning classification approaches (Lu *et al.*, 2007); support vector machines (Webb-Robertson *et al.*, 2008); and the Random Forest classifier technique (Fusaro *et al.*, 2009). These are valuable, because being able to anticipate which peptides will be observable in MS can help experimental design for targeted MS workflows, as explained in the following section.

1.2.3 Proteins can be quantified using MS

Although protein identification alone can provide valuable information about biological systems, there is a limit to the conclusions that can be drawn from qualitative experiments. Instead, knowing *how much* protein is present is usually more valuable. Indeed, recent findings reported by Uhlen using an antibody-based approach (Uhlen and Ponten, 2005) illustrate that differentially expressed proteins are actually rare, because most proteins in humans are expressed in all cell types, most of the time.; less than 1% of proteins are expressed in one tissue only (Service, 2008a). It is precise regulation of protein expression in space and time that results in tissue-specificity; thus, qualitative (present or absent) protein biomarkers are usually less valuable for understanding the true complexities of biology. Indeed, many diseases, including asthma, arthritis, schizophrenia and heart disease, are known to involve very small changes in the regulation of protein expression over time, and diseases are complex, often polygenetic and environmentally-dependent, meaning

that the protein profile for healthy, susceptible and diseased states varies in very subtle ways; hence small changes in quantity may be important factors.

In recent years, several techniques to quantify proteins by mass spectrometry have emerged (Figure 10).

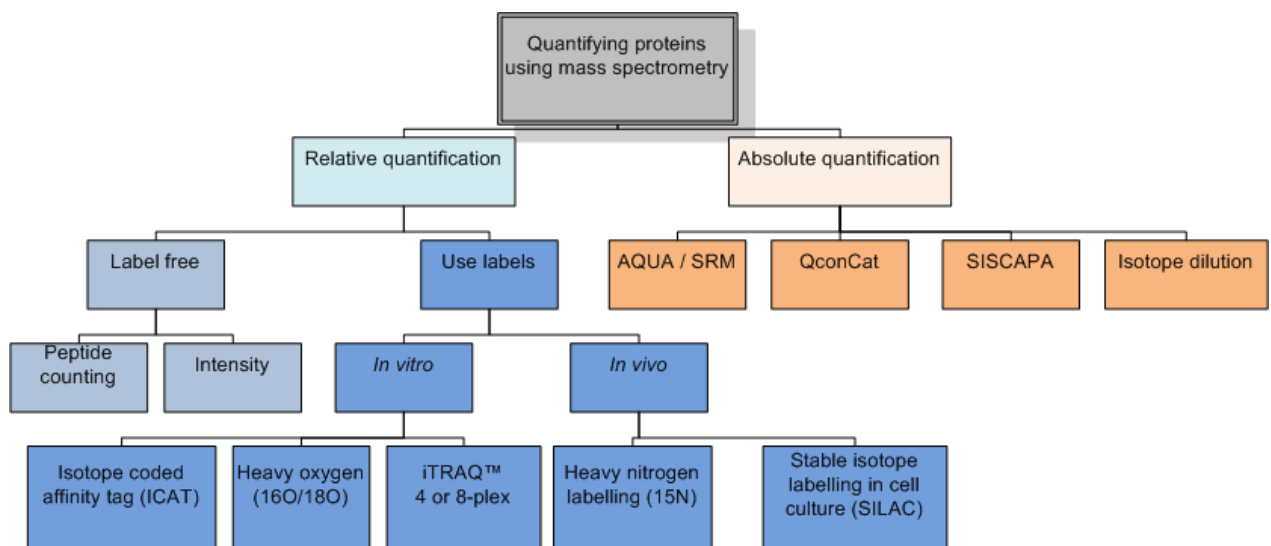


Figure 10 A summary of the approaches for quantifying proteins using mass spectrometry, adapted from (Lau *et al.*, 2007)

Relative approaches measure the abundance of proteins as a ratio between different samples, whereas absolute approaches provide a specific quantity value, such as ng/ml for a specific protein(s).

Relative quantification of proteins (with labelling)

An isotopic label is the marker used to distinguish two (or more) populations of proteins. Heavy elements (heavy nitrogen or oxygen) or heavy amino acid residues

(such as heavy arginine or iso/leucine) are assimilated via cell metabolism using the *in vivo* approach. In stable isotope labelling of amino acids in cell culture (SILAC) (Ong *et al.*, 2002), for example, two cell cultures are grown up in parallel, one exposed to medium containing heavy amino acids, and one with standard medium. For higher organisms *in vitro* labelling with chemical reagents is performed after the proteins have been harvested, such as a label specific for cysteines in the isotope-coded affinity tag (ICAT) approach (Gygi *et al.*, 1999) , and free amine labels with iTRAQ™.

The assumption in all cases is that the proteins will incorporate the labels to completion. Protein quantitation is achieved by comparing the MS intensity of the peptides derived from the two samples, which is possible since the expected mass shift of the ions is known.

Relative quantification of proteins (label free)

Label free methods include peptide/spectral counting and ion intensity monitoring. In spectral counting, protein quantities are estimated in distinct samples by counting the number of MS/MS spectrum-sequence matches found by the search engine. The assumption is that protein abundance is correlated with protein coverage and the number of times a peptide is observed (in replicate experiments). The size of the protein affects the reliability of this approach, and the assumptions are quite 'loose'. For the ion intensity approach, the RP-HPLC retention profile (ion chromatogram) for each peptide is exploited. Chromatographic peak intensities are retrieved and

used as a basis to compare with peptides matched in different experiments. When several peptides are matched to a common protein, each peptide ratio is used to measure protein fold change across the different experiments. Peptide RT can present a problem for this approach; for example, when multiple peptides elute at the same moment peak intensity will be too high, but if peptides are separated in SCX not all the peptides will necessary arrive in the same peak.

Absolute quantification of proteins (SRM and variants)

The AQUA technique (Gerber *et al.*, 2003), now more frequently referred to as selected reaction monitoring (SRM) (Anderson and Hunter, 2006), and QconCat (Beynon *et al.*, 2005) are methods for measure absolute quantities of protein using tandem MS.

To determine the quantity of a specific protein using SRM, QQQ-MS is performed as RP-HPLC separation is in progress. Each peptide is analysed by selection on the basis of m/z using the first quadrupole (Q1). Once separated, in a second quadrupole (Q2), the peptide undergoes fragmentation, generating product ions exclusive to the precursor, which are selectively monitored by a third quadrupole (Q3) (Figure 11). The two stage filtering process in SRM allows chemical background to be overcome by improving signal to noise ratio, and permits several transitions to be monitored quickly. The observed m/z ratio of a precursor peptide and its corresponding product ion is referred to as an SRM 'transition' and has a specific RT associated with it.

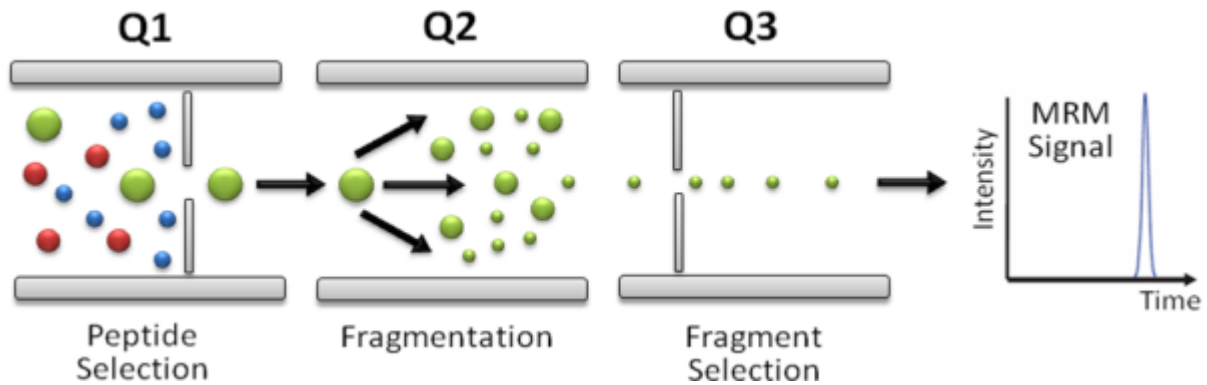


Figure 11 SRM targets a specific peptide to product ion transition in triple quadrupole (QQQ) MS (Source: www.srmatlas.org)

SRM becomes quantitative when the incoming sample is first spiked with a known quantity of stable isotope-labelled synthetic peptide, which is identical in sequence to the expected native peptide (Gerber *et al.*, 2003). For robust studies, calibration curves of serial dilutions of the surrogate are determined to produce more reliable measurement of the quantity of protein.

To monitor a protein of interest, it must be known in advance which transition is most suitable. In simple protein mixtures, a single transition may be sufficient to monitor a particular protein of interest, but in complex samples, such as serum, multiple transitions are generally required due to noise and proteins of very high abundance interfering with the signal (Kay *et al.*, 2007, Keshishian *et al.*, 2007). Furthermore, for very complex samples, such as whole serum, stable isotope standards and capture by anti-peptide antibodies (SISCAPA) (Anderson *et al.*, 2004, Anderson *et al.*, 2009) can be used to enrich the peptide targets prior to SRM to

further improve signal-to-noise. With this approach, antibodies are raised to the native and surrogate (labelled) peptides, and are immobilised on a surface. Prior to SRM, the peptides are enriched by pull down on these antibodies, thus increasing the sensitivity of the SRM assay.

QconCat is a variation on the AQUA/SRM theme, whereby the heavy surrogate peptides are not synthesised artificially, but rather the surrogate peptides are expressed *in vivo* by a bacterial cell expression system. A gene is constructed to encode suitable surrogate peptides (Q-peptides) concatenated in a protein (QCAT protein). This protein is expressed and is digested into peptides for subsequent targeted monitoring (as per SRM). Q-peptides are heavy due to metabolic labelling, where the bacterial cells are grown in culture containing heavy amino acids, for example.

SRM is an increasingly popular technique because it offers the option to measure protein regulation across many targets simultaneously, and in a quantitative manner examples include: (Zhang *et al.*, 2004, Kuhn *et al.*, 2004, Beynon *et al.*, 2005, Unwin *et al.*, 2005, Cox *et al.*, 2005, Ciccimaro *et al.*, 2006, Anderson and Hunter, 2006, Rifai *et al.*, 2006, Wolf-Yadlin *et al.*, 2007, Stahl-Zeng *et al.*, 2007, Kay *et al.*, 2007, Keshishian *et al.*, 2007, Lenz *et al.*, 2007). It is possible to multiplex quantitative measurement of peptides because each transition (the pair of precursor and product ion m/z s) is unique for each peptide. The assumption is made, however, that each SRM does not

interfere with any other in the assay, and that peptide co-elution from HPLC is avoided.

The capability for multiplexing means SRM is often referred to as MRM (multiple reaction monitoring) by practitioners. However, MRM may actually refer to either a set of SRM assays that are being performed for several protein targets simultaneously (the multiplexed approach), or may refer to monitoring multiple product ions (in effect, multiple transitions) for each peptide precursor ion in a single SRM assay. SRM is the accepted MS nomenclature according to IUPAC (Murray *et al.*, 2005), however the term MRM is still widely used in the community, so both are applied in this thesis, depending on the context.

SRM has proven to be a successful method for discovery and validation of novel biomarkers (Kuhn *et al.*, 2004, Zhang *et al.*, 2004, Anderson and Hunter, 2006, Rifai *et al.*, 2006) and, compared to the alternatives, (such as ELISAs) it has the advantage of being cost effective, quicker to design and suitable for multiplexed analysis (Stahl-Zeng *et al.*, 2007). MRM studies have also reported measurements down to attomolar concentration (Onisko *et al.*, 2007, Keshishian *et al.*, 2007). Increased throughput is also possible with SRM, due to direct coupling of separation (via HPLC) to MS and, in some cases, the ability to avoid extensive sample preparation before analysis (Kay *et al.*, 2007, Keshishian *et al.*, 2007). Furthermore, SRM requires a high level of ion separation, but not necessarily high resolution, meaning that the

instrumentation (QQQ-MS) is potentially affordable compared with the alternatives (Orbitrap and FT-ICR); since lower resolution MS is generally less expensive to run.

These features suggest that in the near future SRM may become a routine assay in the clinic. Indeed, transitions for monitoring blood proteins (Anderson and Hunter, 2006, Keshishian *et al.*, 2007, Kay *et al.*, 2007, Stahl-Zeng *et al.*, 2007, McKay *et al.*, 2007) and biomarkers for arthritis (Kuhn *et al.*, 2004), acute liver damage (Zhang *et al.*, 2004) and cardiovascular disease (Anderson and Hunter, 2006) have already been published, no doubt with this objective in mind. In addition, post translational modifications (Cox *et al.*, 2005, Ciccimaro *et al.*, 2006, Stahl-Zeng *et al.*, 2007) and cell signalling networks (Wolf-Yadlin *et al.*, 2007) may also be characterised using the SRM technique.

SRM has only recently been applied to quantify proteins, having originally being used to determine small molecules (Kerns and Di, 2002, Kovarik *et al.*, 2007, Singhal *et al.*, 2007) and metabolites (Gu *et al.*, 2007) in complex sample background by the pharmaceutical industry. This means much has to be learned as regards the optimal approach to designing transitions targeted at proteins: the major challenge being the decision of which peptide(s) are best to monitor, since each protein has multiple tryptic cleavage sites. Moreover, some peptides and their product ions are always visible in MS/MS, whereas others are not detected at all because they do not ionise, for example (Tang *et al.*, 2006).

1.2.4 An introduction to proteomic bioinformatics

In very large volumes of data, the meaning may be hidden. To interpret and understand what large datasets (such as hundreds of spectra) are showing, the data must first be manipulated with computers to make it understandable. An overview of the types of bioinformatics tools and resources for proteomic MS is shown in Figure 12.

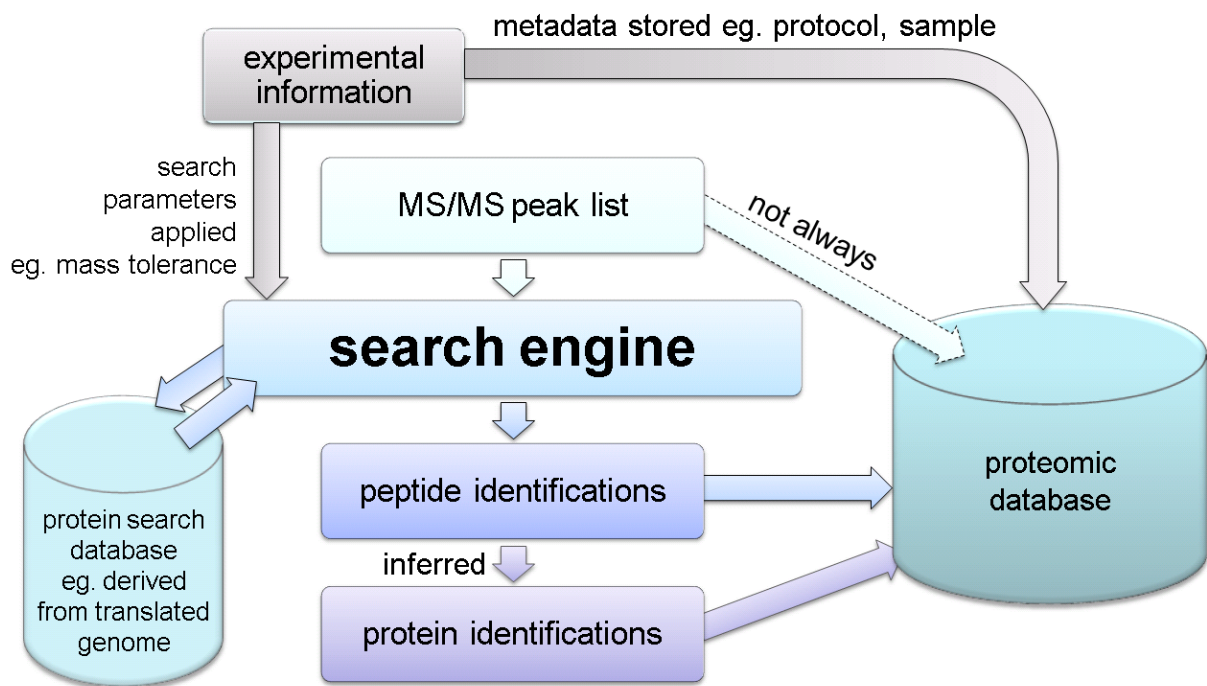


Figure 12 An overview of proteomic bioinformatics resources for tandem MS. Note that ‘proteomics’ repositories are distinct from ‘protein’ databases; the former contains MS/MS data, auxiliary information (such as protocol, species, etc.) and peptide and protein identifications, whereas the latter contains protein sequences, and sometimes often information on protein structure, function, and other properties, for example. (Source: author’s own summary)

Proteomic MS data has 'peak lists'

Molecular ions in mass spectra are referred to as 'peak lists'. This is a list of m/z values (x-axis) and intensities (y-axis). Peak lists are not 'raw data' direct from MS instruments, which is machine-readable and requires proprietary software to convert it to a readable/searchable form, instead they are lists of values either in text file form (such as .mgf, .pkl or .dta format) or encoded in XML (such as mzXML, mzData or mzML) (Figure 13).

```

</contactInfo>rkj@lilly.com</contactInfo>
</contact>
</admin>
<instrument>
<instrumentName>LCQ Deca XP</instrumentName>
<source>
<cvParam cvLabel="psi" accession="" name="type" value="ESI"/>
</source>
<analyzerList count="1">
<analyzer>
<cvParam cvLabel="psi" accession="" name="type" value="PaulTrap"/>
<cvParam cvLabel="psi" accession="" name="resolution" value="2000"/>
<cvParam cvLabel="psi" accession="" name="accuracy" value="0.2"/>
</analyzer>
</analyzerList>
<detector>
<cvParam cvLabel="psi" accession="" name="type" value="EM"/>
</detector>
</instrument>
<dataProcessing>
<software>
<name>PSI-MS XCalibur RAW converter</name>
<version>1.04</version>
</software>
<processingMethod>
<cvParam cvLabel="psi" accession="" name="deisotoped" value="false"/>
<cvParam cvLabel="psi" accession="" name="chargeDeconvolved" value="false"/>
<cvParam cvLabel="psi" accession="" name="peakProcessing" value="centroid"/>
</processingMethod>
</dataProcessing>
</description>
<spectrumList count="2139">
<spectrum id="1">
<acqDesc>
<acqSettings>
<acqSpecification spectrumType="discrete" methodOfCombination="sum" count="1">
<acquisition acqNumber="1"/>
</acqSpecification>
<acqInstrument msLevel="1" mzRangeStart="300.000000" mzRangeStop="1500.000000">
<cvParam cvLabel="psi" accession="" name="type" value="full"/>
<cvParam cvLabel="psi" accession="" name="polarity" value="+"/>
<cvParam cvLabel="psi" accession="" name="time.min" value="0.005833"/>
</acqInstrument>
</acqSettings>
</acqDesc>
<mzArrayBinary>
<data precision="32" endian="little" length="346">
i0W08b/10PMTJhD0+YQ6RVaUPK2pLdgsnaQ2Bom0Pa3Jtdpm+c02z6nENWd51DstedQlxAn0Pm529Dvmag07zwoE0g2qFDem0iQ236okMtr6MDHlekQwppwUMAWa2DsuaQ32op0PABKdDECPqw5LqkMYbatDI0erQzBlrEPGTKlDeg6uQ5RuzkPk565D/HuvQ9LN
)PygLFDihay06yckNc9bJd5GcZQxYUteEM2d7ROBEq1Q+Bntk0yxr2D3Ge3Q2paUE01z7hDoFV5Qz4oukMQelPd/G68Q2I5vYOKQb5DmJm+Q1RXvON8MBEDoMDA09BnuUPqbcJ0ygECQxz3wkP6VMMDFjDFQ23G6EO4d8VdajbH0xDqyEP4RcLD5A/LQz1cyONMhMxD
fjM06agzUM0f85DdnjPQ1bNz00yftEDmG750+wNLEMc3dRD8irYQ6Kt1UMQHNZDKMHwQyjr10M42M3d4izY06au2EPU691DAtn200Lq2kNqL9xDcvbc07BF3UN4zN1D6HPeQ5j3kMWZd9DFJg0+o44UNQFwJD2vbi0ywh5E0vw+RD36P1QwC05UNC4aZd6G/nDxhC

```

(a)

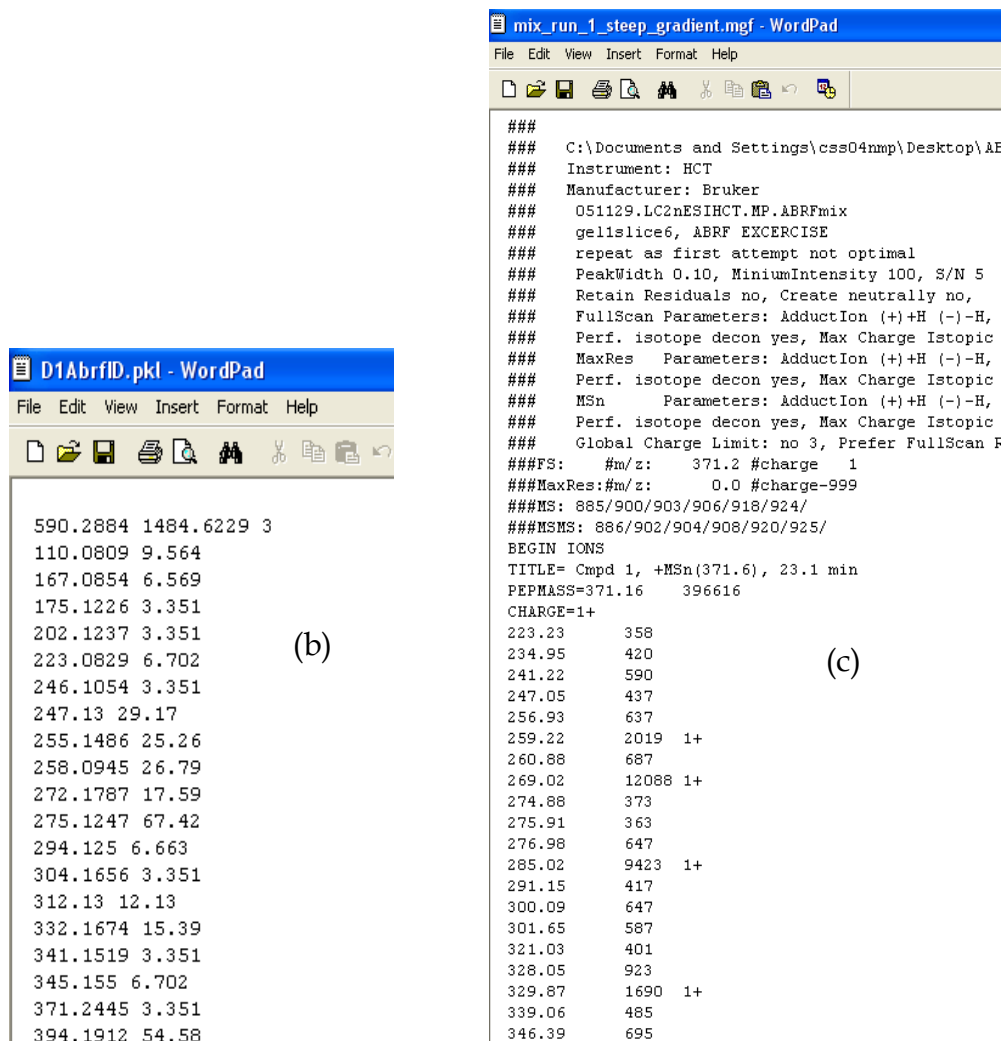


Figure 13 Examples of MS/MS data in peak list form, ready for analysis with a proteomic search engine. (a) mzData is an example of an XML-based format; (b) pkl, and (c) mgf formats are text-based.

Data formats for proteomic MS came about independently and before standard data formats and reporting standards had begun to be established. As a result, formats and repositories were autonomously designed, with no inter-change being possible: the European Bioinformatics Institute developed mzData and PRIDE XML; Ron Beavis' group BIOML and XIAPE; and the Institute of Systems Biology (ISB) mzXML, for example. This remains an issue, because it is often difficult for users to convert their data into a suitable format for other labs to use. However, barriers to

data sharing are beginning to come down as public data format conversion software is becoming increasingly easier to find (see the Appendix I for a summary), and tools have been specifically developed, such as PRIDE Converter (Barsnes *et al.*, 2009) to ease submission to public databases. Academics and MS instrument vendors are also cooperating with the Human Proteome Organisation's Proteomics Standards Initiative (HUPO-PSI) to make their native formats compatible with tools and databases. Furthermore, HUPO-PSI is coordinating the development of standard formats for the major proteomics data-types (Table 3); for example, recently competing MS data formats mzData and mzXML have been replaced by a single new format called mzML (first release June 2008), which is expected to become the universal unprocessed proteomic MS data format. Furthermore, MRM transitions now have TraML format.

Table 3 Standard formats for proteomic MS developed by PSI. Protein separation, interactions, and modifications are managed by other groups and their formats.

Standards Work Group	Format name	Format remit
Mass Spectrometry	mzML	Merges various peak list formats (not native .raw formats). It has experimental information (metadata). Peak lists are encoded in binary to make files manageable in size
Mass Spectrometry	TraML	MRM transitions (see Chapter 6 for more information)
Proteomics Informatics	mzIdentML (formerly analysisXML)	Peptide/protein identifications from search engines plus search metadata, decoy database used, etc.

There are also work groups for developing standard formats for protein modifications, interactions and protein separation.

Metadata for proteomic MS

Data that describes the MS/MS peak lists is called 'metadata'. Metadata is usually required to identify proteins using an automated system (such as a search engine). Metadata for this purpose includes the protein search database, the mass tolerance to be applied in the search (which represents the level of mass accuracy achieved by the mass spectrometer at MS and MS/MS levels), the proteolytic enzyme used, any anticipated PTMs, and the expected charge state of the peptides in the experiment (usually 2+). An example of a metadata entry form is in Figure 14.

Data Profile ↓

Profile Name: (max 20 letters)

Search Property: Public Private Hidden

Cell line type: (or blank if not cell line)

Cell Type:

Development stage:

Sample collection time:

Tissue Types: (Please fill in if tissue is not in list)

Disease States: (Please fill in if disease is not in list)

Fraction/depletion info:

Sample Species: (Please fill in when you choose 'other')

MASS options: Monoisotopic Average

Search dB:

Digestion enzyme:

Allow up to missed cleavages

Fixed modifications:

Variable modifications:

MS tol.± Da

MS/MS tol.± Da

Peptide charge up to:

Instrument

Instrument Calibration

Figure 14 The metadata required by the Genome Annotating Proteomic Pipeline, which employs the X!Tandem search engine and APS

This form also includes ‘biological’ metadata. These are descriptions of the experiment, such as the name of the instrument, the cell line or tissue, disease state,

protocol details, such as the separation protocol used. Having this information available means that the resulting identifications, when stored in a database, can be data-mined to reveal hidden trends. For example data-mining can reveal the proteins present in certain disease states, or show that certain proteins are present under particular processing protocols and absent in others.

Peptide search engines

Proteomic search engines are central to the field of proteomic bioinformatics. There are both free (X!Tandem (Craig and Beavis, 2003)) and commercial (Mascot (Perkins *et al.*, 1999)) offerings, with each being based on differing scoring algorithms. In proteomics 'discovery' studies, the protein content of the sample being analysed is usually unknown, however, there is still some information that can be utilised to assign sequence to m/z . It is known, for example, that all proteins consist of combinations of only 20 possible amino acid units, each having a known molecular weight. It is also known that trypsin, the enzyme most often used, cuts proteins after lysine or arginine (if not followed by proline). Most importantly, the target protein sequences are known, as database, such as a translation of the genome; thus information from the genome can be applied to characterise the proteome using search engines. In summary, using all these sources of information, combined with statistical methods to assess the likelihood of matches compared to chance, various search-engine type algorithms have been designed. Most search engines fall into four types of algorithm:

1. **Peptide Mass Fingerprinting (PMF)** matches masses (as m/z) to peptides (Figure 15) by comparing the unknown masses to a theoretical database, generated by performing an *in silico* cleavage operation (per trypsin) on a protein sequence database. With PMF, several peptides may share the same mass because permutations in the arrangement of the same set of amino acids will result in peptides of identical mass. And certain combinations of amino acids may result in mass differences, which are indistinguishable at the resolution of MS. Once masses have been assigned to peptides, proteins may be identified as those that contain a number of the matched peptides.

2. **Tandem MS searching** is the most prevalent approach. As with PMF, the unknown spectrum is compared to a theoretical database, but this time the database is derived from *in silico* digestion of proteins, followed by theoretical CID fragmentation of the peptides (Figure 15). Product ion masses are used, so it overcomes the problem seen in PMF, allowing peptides to be distinguished even if they have identical mass.

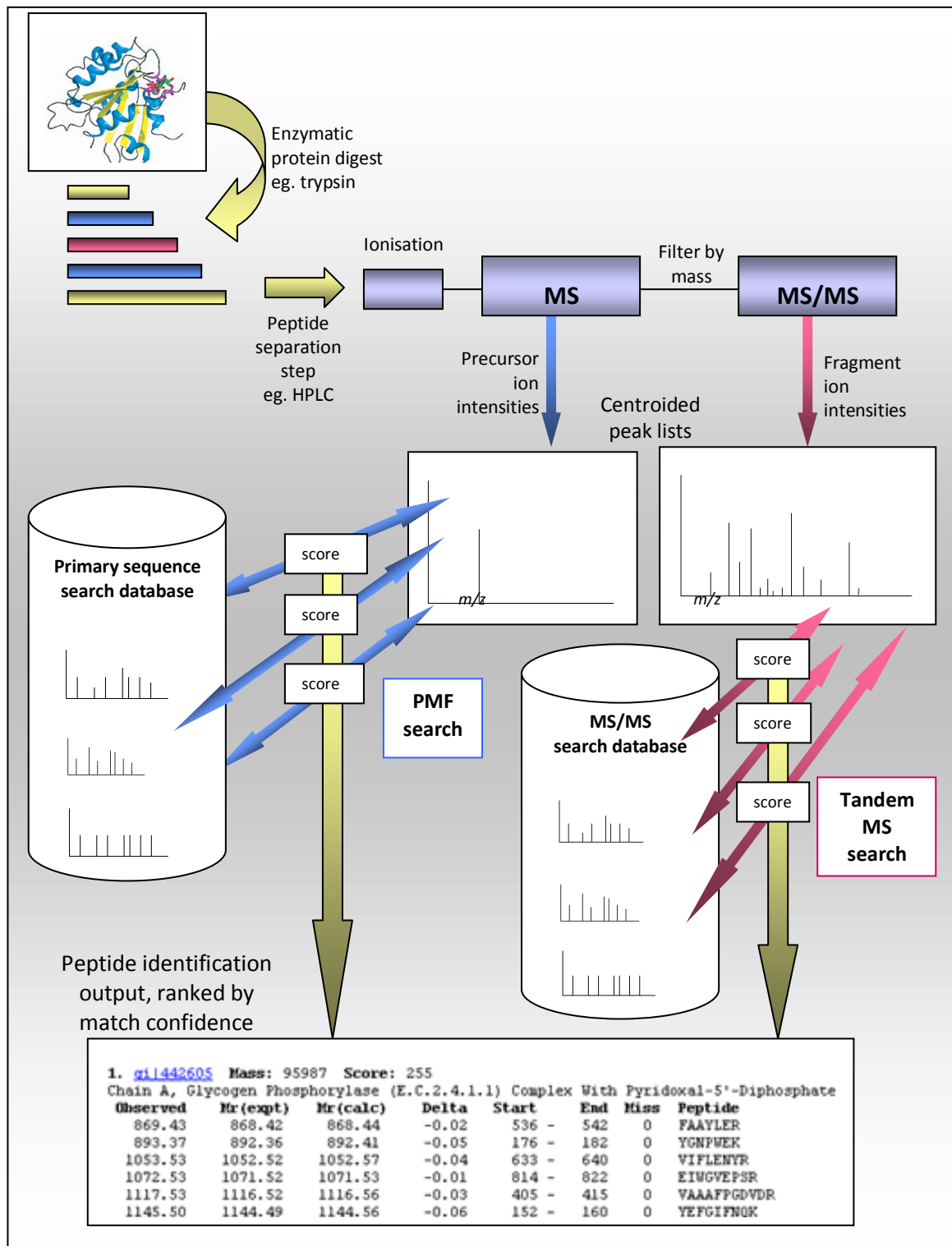


Figure 15 The principle of spectrum identification using PMF and tandem MS searches. Scores are generated by comparison of experimental peak lists with theoretical peak lists in the search databases. For PMF, the theoretical fingerprint, which is the unique set of peptide masses generated by *in silico* enzymatic cleavage of a protein or translated genomic database, is matched to a single pass mass spectrum. For tandem MS searching, the search proceeds initially in a similar way to PMF, by matching to the precursor ion mass, but following this, the experimental tandem spectrum is compared in a second round of searching to a theoretical peptide fragment spectrum of daughter ions generated for each candidate peptide, based on known peptide fragmentation chemistries. It is possible to identify proteins by assigning MS/MS fragments only. Note that all searches are based only on the m/z component of the spectra, signal intensity is only used to discriminate peaks from noise. (Source: author's own summary)

3. **De Novo sequencing** seeks to use the tandem MS spectrum as the sole reference for deducing the sequence of the peptide it represents, it does not apply a search database, only knowledge of amino acids and CID fragmentation.





As product ions are generated by splitting amino acids, all the information necessary to reconstruct a peptide sequence can exist in the ions generated, although spectra of high enough quality are in reality quite rare. This technique can be performed by hand and is advantageous, for example, when no reference database is available, as in the case of organisms with an as yet unsequenced genome.




4. **Peptide sequence tagging** locates a peptide sequence by searching a database with partial sequence information, termed 'tags', derived from the spectrum. An example of a sequence tag is '[340]GLGSA[112] PK', where the letters denote amino acid sequence which could successfully be identified from the spectrum using *de novo* methods, and the numbers in brackets represent unknown amino acid combinations with mass equal to the values (because sequence could not be established from the spectrum alone for these regions). The tag is then used to query the search database for possible matches.

Development of peptide identification programs is a major area of bioinformatics research. Table 4 summarises some popular search engines in routine use, and gives a synopsis of the algorithms they employ. Due to the open source philosophy of the

bioinformatics research community, many tools are freely available to use and download; review articles provide a comprehensive overview (Sadygov *et al.*, 2004, Xu and Ma, 2006, Shadforth *et al.*, 2005a).

Table 4 A selection of the major protein/peptide identification search engines

Program	Ref.	Type of search algorithm	Description	Free to use on web	Free to download	Website
	(Pappin <i>et al.</i> , 1993)	PMF, tandem and tagging	Uses the MOWSE algorithm. Calculates the probability that the match was observed by chance, and by using knowledge of the exact size of the search database, calculates the statistical significance of the match. Some additional heuristics for intensity and ion ladders are included in the search.	Yes	No	www.matrixscience.com/
	(Eng <i>et al.</i> , 1994)	Tandem	The protein database is searched to identify linear amino acid sequences. A cross-correlation function is then used to provide a measurement of similarity between the theoretical mass for the fragment ions in the database and the real mass of fragment ions observed in the tandem mass spectrum.	No	No	Sequest sourcerer and cluster are Thermo Fischer products. www.thermo.com
	(Craig and Beavis, 2003)	Tandem	'Dot product' method is used correlate theoretical spectra with the real one. Performs multiple stages of searching and refinement to ensure efficient matching of mass peak lists to sequences, and to optimise for speed. The first step aims to match the theoretical tryptic peptide and fragment masses to the real MS signal peak lists. Then further iterative steps search for PTMs and point mutations, thus search space is decreased to a manageable size for computation, and more of the peaks are successfully assigned.	Yes	Yes	www.thegpm.org/TANDEM/index.html
	(Colinge <i>et al.</i> , 2003)	Tandem	OLAV algorithm. Stochastic model, the parameters for which are learnt from a set of reference matches. The likelihood of the match is derived from deciding between the null hypothesis, that the match is random, versus the alternative, that it is correct. Also can identify known PTMs in the search.	Yes	No	www.phenyx-ms.com/

	(Ma <i>et al.</i> , 2003)	De novo and tagging	Determines corresponding peptide sequences without the use of a theoretical peptide fragment database. All possible amino acid combinations are calculated and the peak lists are matched to these.	No, but free web demo	No, but free demo	www.bioinformaticsolutions.com/products/peaks/
GutenTag	(Tabb <i>et al.</i> , 2003)	Tagging	Uses an empirically-derived model of fragment ion intensities to increase accuracy when deriving sequence tags, which are used to search the database. Score of the match is determined by comparing the experimental spectrum and theoretical spectrum using the model.	No	No, license required free for academic users	http://fields.scripps.edu/GutenTag/
	(Geer <i>et al.</i> , 2004)	Tandem	The Open Mass Spectrometry Search Algorithm - scores significant hits with a probability score developed using classical hypothesis testing, the same statistical method used in BLAST.	Yes	Yes	http://pubchem.ncbi.nlm.nih.gov/omssa/
	(Hummel <i>et al.</i> , 2007)	Tandem	'Modified' dot product distance measure between unknown and reference spectra. MS/MS spectra (unknown peak list, x, and library of peak lists, y) are compared and a distance between them is found using a series of steps. The distance is calculated and taken as a single overall measure of the goodness of match between unknown and library. To determine which distances are true matches a threshold is applied to the distances.	Yes	No	http://www.promexdb.org

Metadata is important for the accuracy of search engine results. For example, if the mass tolerance window is set too large, ambiguous peptide matches may result leading to false peptide matches. Other considerations are search space, for example, if many PTMs are specified the search may take a long time. Having accurate metadata is important for the search accuracy, but also for users wishing to interrogate the resulting identifications later.

Harnessing search engines to make large scale sets of identifications is powerful and widely used. Nevertheless, despite recent advances only 10-20% of the MS/MS spectra analysed can be confidently assigned to peptide sequences. Reasons for uncertainty come from many sources, for example: trypsin cleavage may not proceed to completion (depending on the protocol used); there can be inadequate mass resolution by the MS instrument to be certain of identity; and the search space can be too large to explore especially when the various possible protein modifications (such as oxidation or phosphorylation) are included in the search. Recent work by Matthias Mann's group is improving the situation, whereby up to 90% of all fragmented peptides in yeast could be identified using high resolution Orbitrap (linear ion trap) instruments and a new search algorithm, MaxQuant (Cox and Mann, 2008). In this case, the number of assignments was boosted by the 'robust scoring' applied to peptides that were found to be modified versions of already assigned peptides. This approach of re-searching for modified variants is applied in X!Tandem, and GAPP pipeline.

Proteomics search databases

The search database applied for the peptide identification process is important. Increasingly, consensus spectra databases are being used for 'pattern matching'-type searches, instead of theoretical databases searches. A consensus spectrum is a fusion of all available experimental MS/MS spectra for an individual peptide sequence into a single composite spectrum, which retains the most frequent m/z and intensity features. The consensus spectra are also usually MS instrument-specific, since each approach favours detection of different fragment ion types. The search process is analogous to methods used for small molecule identification, such as in metabolomics (Dunn and Ellis, 2005), whereby annotated spectrum libraries are searched for matches with unknown sample spectra.

Since more MS/MS spectral data is available than ever before, there is now scope to construct this kind of database, whereas it was not an option before due to the lack of data. X!Hunter (Craig *et al.*, 2006) and SpectraST (Lam *et al.*, 2007) are search engines designed to use consensus spectra for identifying peptides. These are the first examples, but the approach will no doubt increase in popularity as more data becomes available. This is because, compared to searching a theoretical spectral database, consensus spectra offer faster identification speeds. Moreover, by comparing real experimental (unknown) spectra to real (known) spectra, it increases the chances of correct identifications, because phenomena experienced in the real spectra will be captured in the search. Theoretical spectra, on the other hand, may

capture the stochastic element apparent in real spectra, so may still have value in assisting identification.

For the purpose of comparing the identifications derived from search engines (only PMF, tandem, tagging searches and limited support for *de novo* sequencing) there is a new standard data format, mzIdentML (formerly known as analysisXML) being developed by HUPO-PSI. This XML-based format provides sufficient information to enable a subsequent researcher to run the same search using the same or another search engine, permitting validation of results by other scientists or reviewers. It also supports enough information for tools to display the spectral evidence (if available) to demonstrate the peptide matches, including support for isotope labelling studies. Metadata (search parameters), manual protein annotations, as well as the search database used are captured (but not the peak lists *per se*, only a reference to them).

Inferring proteins from peptides

All search engines generate peptide identifications. Protein identifications, however, are made subsequently by inferring a match, given the peptides found. In most cases, users first validate the peptides found; options to do this are in Figure 16.

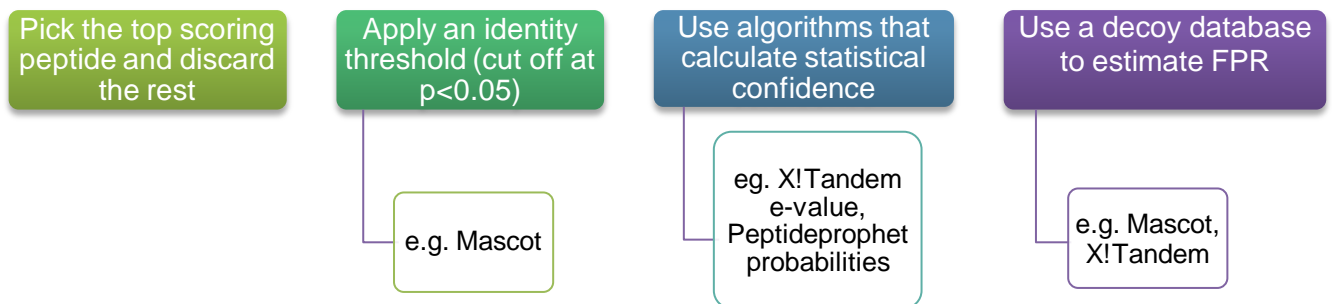


Figure 16 Peptide assignments from search engines may be filtered by user-specified criteria to attempt to remove false identifications before protein inference begins

There are various algorithmic approaches to infer proteins from peptide identifications. The principle of parsimony is sometimes used to overcome the problem of ambiguous peptides (Figure 17).

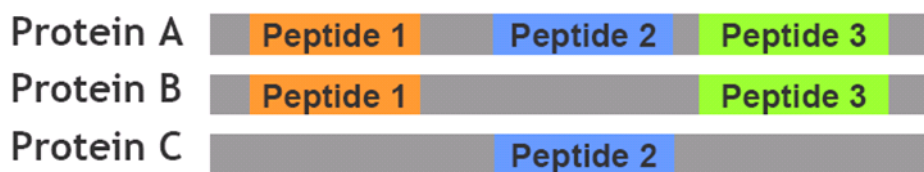


Figure 17 The problem of protein inference. The principle of parsimony (Occam's Razor) favours Protein A.

Indeed, if the protein molecular weight and pI of the proteins are available (from the 2D gel stage), then protein identification is made easier, however, this is usually not the case in high-throughput setups. A tutorial on protein inference (Nesvizhskii and

Aebersold, 2005) suggests that peptides should be distributed using a probability-based approach that takes the probabilities of peptide assignments into account (Nesvizhskii *et al.*, 2003b). This has an advantage of permitting calculation of statistical confidence measures for the protein identifications and allows estimation of false identification error rates resulting from filtering the data. Indeed, Proteomics journal insists that for each protein identified there is a stated measure of certainty, such as a p -value.

The identification of a single peptide is not usually deemed acceptable to confirm the presence of a protein; these are seen as unreliable thus are usually unacceptable for publication. The more peptides are matched the more likely it is that the match is correct, and the protein isoform (splice variant) can be confirmed.

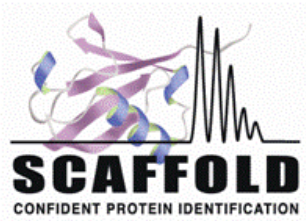

Validating protein identifications

The problem associated with automated searches is finding a means to assess the accuracy and hence reliability of the results, because without human interpretation false positive (FP) identifications may not be recognised as such. FPs come about for several reasons, including unexpected enzymatic cleavage, lack of sufficient mass resolution, poor sample handling and background noise. The cost of FPs can be substantial because such identifications can feed through into later stages of research, leading to, for example, phantom biomarkers or erroneous drug targets or erroneous conclusions about the underlying biology.

'Post-identification systems' are available to perform validation of protein identifications made by automated searches (Table 5). In some cases, the validation and protein inference are performed by the same program as for APS in the GAPP pipeline, and ProteinProphet, for example.

One approach to confirm identifications are correct is by taking a consensus across different search engines and deriving a composite score based on performance across the board (Alves *et al.*, 2008, Jones *et al.*, 2008a). Consensus comparisons may be performed by Scaffold software (a proprietary offering), for example (Table 5).

Table 5 Post identification systems validate identifications derived from automated searches

Program	Reference	Type of system	Description	Free to use on web	Free software download	Website
	(Searle <i>et al.</i> , 2008)	Validation of diverse search engine results to reduce false positives	Performs multiple searches on the same <i>raw dataset</i> in parallel and thereby compares independent interpretations of the same data, providing a confidence score of the combined results	Yes	No	www.proteomesoftware.com/
	PeptideProphet (Keller <i>et al.</i> , 2002a), ProteinProphet (Nesvizhskii <i>et al.</i> , 2003b)	Validation of various search engine results by using probabilistic methods. Enhances existing search engine identification outputs by permitting comparison – with probabilities being on the same scale.	PeptideProphet analysis is followed by ProteinProphet, which groups peptides by their corresponding protein(s) to compute probabilities that those proteins were present in the original sample.	Yes	Yes	http://peptideprophet.sourceforge.net/

**Advanced
Average
Peptide
Score (APS)**

APS
(Chepanoske *et al.*, 2005),
advanced APS
(Shadforth *et al.*, 2005b)

Applies a threshold
and, protein
inference

False positive protein
identifications may achieve high
scores as a result of many low-
scoring peptides. The average of
peptide scores (APS) is
calculated and any proteins
with an APS lower than a
certain threshold is deemed
false. The threshold is set by a
decoy DB search, which is used
to filter the hits returned from
the target DB search (explained
later).

Yes, as
part of
GAPP

No

www.gapp.info

Another approach to reduce FPs is decoy database searching. Using a decoy database it is assumed that the highest scoring match (between a peptide and a spectrum found) in the decoy search is a suitable threshold for filtering off the incorrect identifications found in the target search, and hence is used to eliminate FPs. In this way, the decoy is exploited to test the null hypothesis; this being the test to determine whether a peptide was not identified in MS. The caveats with this include the time consumed in searching the same dataset twice, and the difficulties in creating a mirrored distribution of the target without including any target sequences - to avoid forfeiting sensitivity. Methods to reduce FP detection by decoy database searching have been published recently, including modification of PeptideProphet to incorporate a target-reverse proteome search (Choi and Nesvizhskii, 2008b), incorporation of a decoy database search option into Mascot and research articles such as (Elias and Gygi, 2007, Reidegeld *et al.*, 2008).

Proteomic pipelines

Proteomic pipelines are the multi-step processing platforms required to seamlessly convert tandem MS/MS data into protein identifications by calling a specific search engine. Using pipelines, thousands of spectra may be processed consistently and in manageable timescales; batch spectra submissions are usually possible. In this way, pipelines provide a convenient route into proteomics repositories. Pipeline steps can include: data format conversion, quality filtering, execution of a search engine and post-processing validation. Examples include Global Proteome Machine (GPM), Trans-Proteomic Pipeline (TPP), Genome Annotating Proteomic Pipeline (GAPP)

and others. GAPP pipeline is applied to novel research in Chapter 4, and is also the basis of data analysis and capture for the tool developed in Chapter 5. A detailed description of the most frequently used pipelines is presented in Chapter 3.

In addition to existing pipelines, developers may use software frameworks to design their own bespoke, flexible workflows. OpenMS (Sturm *et al.*, 2008), for example, has a static core for download, and via this core OpenMS-specific or external software tools may be called (via the command line) to process HPLC-MS/MS data. There is a library of free packages comprising 350 classes and 100,000s of lines of C++ code. The OpenMS Proteomic Pipeline (TOPP) (Kohlbacher *et al.*, 2007) is one example of an OpenMS pipeline which is ready-to-use and designed to be run locally.

Proteomic MS repositories

Proteomics repositories are for storing, integrating and sharing MS/MS data. Once the peptides and proteins have been identified in an experiment, their usefulness does not stop: by sharing and comparing the protein status of various samples, understanding of biology can be uncovered, for example, the proteins specific to certain tissues or organs (i. e. specific proteomes) can be identified by comparative studies. There are many publicly accessible repositories on the internet, most with additional tools to make understanding the data in the repositories easier; effectively converting meaningless spectra and identifications into knowledge. For example, the PRoteomics IDentifications database (PRIDE) provides a facility to generate

Venn diagrams to compare protein expression profiles between experiments in the database. Moreover, graphical representation of the data, including clickable image maps of tissues where proteins have been identified are available (at the Max-Planck Unified Proteome Database (MAPU), for example), and dynamically generated spectrum views and colour shading to show significance of matches (at GPM database). Programmatic querying (with API) is also available with some repositories, via Biomart, for example.

1.3 Aims and objectives of this thesis

This thesis is focused on key areas of unmet need in proteomic bioinformatics research and management. For the management component, this EngD aims to establish the current state of exploitation of proteomic bioinformatics in public and commercial environments in the UK and abroad. It looks at why the sponsor, GlaxoSmithKline (GSK), downsized proteomics and proteome bioinformatics research, and makes recommendations for future investment in proteomic bioinformatics products and services.

The engineering research component aims to deliver computational solutions to two problems: first, to improve confidence when identifying proteins in MS/MS data, and second, to provide integrated online resources to support the next generation of targeted, quantitative proteomics research. These deliverables will assist proteomics researchers in the community, such as those performing biomarker discovery and validation studies, and will help to make proteomics research cheaper and more

efficient, for example, by reducing the risk of false leads and saving time designing experiments. Moreover, by improving the quality of identification and quantification of proteins, this EngD contributes to the field by moving resources a step closer towards achieving accurate proteome information for modelling the workings of biological systems in the future.

The thesis is split into three parts: Part I (Chapter 2 and 3), Part II (Chapter 4), and Part III (Chapter 5 and 6). The field of 'proteomic bioinformatics' in industry and academia is mapped out in Part I. In Chapter 2, for example, the question of whether proteomic bioinformatics is a commercially viable activity is explored. The author investigates how high-throughput proteomics was funded, focusing on events from 1985-2009. This is an important story that may present new lessons for the funding and development of new high-tech businesses in the future, and although there are rare examples of management literature for proteomics, such as (Mitchell, 2003) and a chapter in (Moody, 2004), there are no examples of investigations into the business history of *proteomic bioinformatics*, so the study will be unique, timely and a valuable resource for management decision-makers and investors in this field.

In Chapter 3, a review of the freely available proteomics data repositories is performed. This review highlights key data sources, which form the foundation for subsequent research in this thesis. Moreover, the review represents novel research,

because a comprehensive overview of public repositories was not previously available.

In Part II, approaches to improve confidence in the results from an automated proteomics pipeline are investigated. Expert judgement calls cannot be made to spot incorrect identifications on a spectrum-by-spectrum basis using a pipeline. Therefore, decoy database searches may be performed to isolate incorrect identifications and filter them out. In Chapter 4, therefore, the author investigates which decoy database design is most efficient at reducing FPR using the Genome Annotating Proteomic Pipeline (GAPP). This is important research, because the cost of pharmaceutical R&D is rising, and by reducing FPRs the risk of pursuing false leads is lessened.

Finally, new computational systems to support the design of quantitative MRM experiments are developed in Part III. For example, designing transitions for MRM *ab initio* is challenging - it often requires empirical 'discovery' studies and expert knowledge - therefore, a new algorithm is developed in Chapter 5 to speed up the design process. The tool predicts the best candidates by leveraging existing public data resources and tools, and combines these with rules captured from expert practitioners. Also, once a transition has been developed and validated in MS it may be reused, hence Chapter 6 describes the development of a new database management system for disseminating validated transitions, so researchers can easily re-use transitions instead of designing them from scratch, and spend less time

scanning heterogeneous literature for a suitable candidate. In turn, this compendium will serve as a 'shop window' to boost exposure of the data submitters' work and increase citations.

The Business History of Proteomic Bioinformatics (1985-2009)

2.1 Summary

In this chapter, the author investigates the emergence of proteomic bioinformatics, the subject of this EngD thesis, and looks at the technological advancements and public and private funding mechanisms that enabled its development into a new industry of its own. The management research performed demonstrates how business and science interacted to create new breeds of high-tech organisations – first, proteomics biotech’s, followed by in-house departments in research and development organisations, and more recently niche firms focused on proteomic bioinformatics. By presenting insights into how this new and specialised industry of proteomic bioinformatics developed, including detailed analysis of companies and operations in “big pharma”⁶, predictions are made for the future model of the proteomic bioinformatics market. These recommendations are aimed at the management decision-makers at the project sponsor (the blue chip pharmaceutical company, GlaxoSmithKline), at investors in biotechs, and generally at decision-makers and stakeholders who may be interested in exploring proteomic bioinformatics activities in the near future.

⁶ ‘Big pharma’ generally refers to large pharmaceutical companies that have political influence. Specific definitions vary but it is a widely used term in management literature. Definitions include, for example, that revenue should be in excess of \$2-3 billion, R&D expenditure should be in excess of \$500 million, and/or the firm should operate in the major global markets (i.e. USA, Europe and Japan).

2.2 Introduction

This chapter describes the development of 'high-throughput' proteomics, and takes the reader from emergence of the science, and subsequent uptake by biotech and pharmaceutical companies, to the current status of the field. It focuses on the interaction between the new science (proteomics and proteomic bioinformatics), commercial markets and public funding bodies. As such, the resulting narrative is referred to as a 'business history' of proteomic bioinformatics.

2.2.1 The proteomic bioinformatics market is investigated using a business history approach

A 'business history' is a story about industry in the past, and can include the history of an individual firm, or entire business systems; often described using a specific time period and geographical location (Amatori and Jones, 2003). The story usually includes the relationships between businesses and their political, cultural, institutional, social and economic contexts⁷. Business history began as a discrete discipline at the Harvard Business School in the interwar years, with the first histories emerging in the 1950s⁸. The business history approach is now widely used, and can contribute to decision-making and strategic management processes, often being published in journals, such as *Enterprise and Society: The International Journal of Business History*; *Business History Review*; and the *Journal of Economic History*. Teaching case studies on an MBA course at Cranfield University, for example, are a type of brief business history, usually focused on a specific firm; and

⁷ Taken from the remit for *Enterprise and Society: The International Journal of Business History*

⁸ Examples include 'The History of Unilever: volumes 1 and 2' (C. Wilson, 1954, London) and 'Pioneering in Big Business' (about Standard Oil) (R.W. Hidy and M. E. Hidy, 1995, New York).

can be a convincing way to demonstrate business principles, stimulate ideas and teach strategy.

The business history and analysis of the proteomic bioinformatics industry presented here is timely for three main reasons. Firstly, decision-makers in big pharma, venture capitalists (VCs) and universities currently have no detailed reports regarding the business of proteomic bioinformatics and its associated technologies. This industry report leads the reader from no knowledge of proteomics to understanding the funding mechanisms, current market and wider context of proteomic science and informatics. With this knowledge, better decisions may be made for funding proteomic bioinformatics and/or other similar technologies in the future. Furthermore, the analysis presented has the advantage of being performed by an author actively involved in proteomic bioinformatics research, with relevant industry contacts and data sources to draw primary data.

The second reason is that during the project (in 2005-6) GSK downsized high-throughput proteomics in R&D. This meant that no suitable new data were generated by the sponsor and hence the output of the project was not a system(s) to specifically meet the sponsor's needs, but rather relied upon publicly-available tools and datasets that were freely-available or from collaborators. For the sponsor, therefore, this business research delivers a new account of the events that lead to the downsizing of proteomics as well as a description of the wider context in which

these events took place. A new, independent perspective on these events is timely and valuable for the community to understand and learn from what happened.

Finally, on a practical level, the business history and case-study-approach are employed, because empirical data and observations are the most accessible source of information on proteomic bioinformatics, which is a specialised, high-tech industry. It is not represented in management literature, so any other approach would not have been possible. Moreover, the approach is a powerful one because by tracing developments over time, looking at the decisions made, funding provided and the players in the market, a picture of the business of proteomic bioinformatics may be drawn, and conclusions offered based on real evidence, rather than referring to the observations of others.

The author believes that in spite of the downsizing that occurred at the sponsor, this management report (and the remainder of the thesis) demonstrates that the tools and techniques developed are timely and valuable for the next generation of targeted, quantitative proteomics research.

2.2.2 Management hypothesis and contribution to knowledge

The main hypothesis for this chapter is that there are links between the economics of proteomic/ proteomic bioinformatics and the way the science *per se* has developed. The main precept is that economic forces pushed the growth of the uptake of the science, not the true ability for the science to deliver valuable products and services.

This is the reason for the wave of investment suddenly drying up, once it was realised that commercially valuable protein biomarkers were not visible on the horizon.

There are two main areas to investigate to demonstrate the truth of this hypothesis. Firstly, did the market for proteomic bioinformatics come about because of a process innovation (such as computer technology, new reagent development)? Or, did it come about because of the excitement and the success of the genome sequencing project? Was it scientific leaps or hype that would fuel growth and development of a new market for proteomic bioinformatics? The business history presented here is a detailed examination of events, which will make the answer to this hypothesis clear.

The contributions to knowledge of this chapter are threefold: this is the first business history ever to be written on the proteomic bioinformatics industry. There are examples of business histories of new scientific fields, such as genetic engineering (McKelvey, 2000), and a review of proteomics as part of the digital revolution in biology (a chapter in (Moody, 2004)), however an investigation into the economics of proteomics/ proteomic bioinformatics *per se* is novel. Secondly, the author has examined the two aforementioned research questions, and makes conclusions based on new history and case-based evidence collated specifically for this purpose; these data have not been described before. Finally, this chapter represents the first attempt to put the business of proteomic bioinformatics into the context of the pharmaceutical R&D industry, and offers new recommendations about where

proteomic bioinformatics in this industry is going next; these predictions are new and unique.

2.3 The business history of proteomic bioinformatics

Events in the following business history are split into phases, and the funding for each stage of development is described. This section serves as preliminary evidence for the recommendations made for potential investors in proteomic bioinformatics in the discussion section.

2.3.1 The early innovators of proteomics and proteomic bioinformatics were publicly-funded

It is commonly thought that proteomics came about as a result of the sequencing of the human genome, which took place during the late 1990s by the publicly-funded Human Genome Project and a parallel effort by Celera Genomics' J. Craig Venter. This is correct, since the now commonly used high-throughput approaches in MS-based proteomics rely on search engines to identify peptides in mass spectra by searching a protein sequence database. To do this, a genome sequence that is translated into proteins is needed, so the availability of a genome was important. However, the philosophy and desire to characterise the proteome *in toto* actually began years before. According to Nicolas Wade⁹ (Wade, 1981) it began with the American pioneer, Norman Anderson, who first attempted to set up a human protein index project as a national objective in 1980 (reference in (Moody, 2004, Fong, 2009)). He failed at that time since gene-related research was the focus for funding

⁹ New York Times science journalist and author

agencies, however, in spite of this he and his son (Leigh) set up the first 'industrial-scale' protein cataloguing project in 1985. With \$1 million of US public funding and a team of around a dozen scientists, they began to create a database of protein maps. This database would contain x,y-coordinates of proteins measured on 2D-PAGE¹⁰ gels; a technique that separates proteins on a square 'canvas' of jelly. The Andersons aimed to quantify each protein by measuring the size of its spot on the gel. This was revolutionary, because it applied Henry Ford's conveyor-belt 'brute force' to biology. The aim was ambitious: involving creation of a catalogue of all human proteins and building a huge virtual repository to store the findings. At this time, Edman degradation¹¹ - a low-throughput chemical technique - was used to identify proteins. This meant the project was colossal in terms of the time and resources to complete it; the Andersons predicted the eventual cost of their protein index to be \$350 million over a further five years. No one had seen protein research on this scale before. Quickly it became clear, however, that they were failing to create a valuable catalogue. The variability of the 2D gel approach meant that each time a gel was run the spots moved, making the database useless for other researchers. Moreover, obtaining suitable quantities of interesting proteins was an issue, since unlike genomics, where PCR can be applied to amplify DNA, there is no such method to increase the quantity of protein. In spite of the obstacles, the Andersons new 'list-based' approach to biology was revolutionary and ambitious. They were early innovators, because their ambitious approach was to characterise the proteomics

¹⁰ See introduction for a detailed description of this technique.

¹¹ Edman degradation, named after its inventor in the 1950s, is a method of determining the order of amino acids in a peptide. It is a relatively time-consuming, chemical method that pre-dates mass spectrometry for peptide sequencing.

biotech entrepreneurs a decade later. Their work provided the first glimpse of the huge hype, and correspondingly large investment that would be stimulated by the promise of proteomics.

Irrespective of the lack of continued funding the Andersons set up a company called the Large Scale Proteomics Corporation (LSPC) to create a proprietary proteomic database using their own gel platform 'ProGEX', which could analyse "1 million proteins a week"¹². In addition to \$1.4 million in internal funds in 1994 they received \$1.9 million of US government's Advanced Technology Program: "*Since no company...had improved ... electrophoresis since its development in 1975, sources of private funding for LSPC's efforts were difficult to find.*"¹². The science was too young, and hence too risky, for private investors to get involved. Several years later (1999), however, their technology was acquired by the Large Scale Biology Corporation (LSBC) and LSBC went on to release the Human Protein Index based on the Anderson's work.

In Europe proteomics was also beginning to attract interest. Oxford Oligosaccharides later to change its name to Oxford GlycoSystems then Oxford GlycoSciences spun out of the Glycobiology Unit at Oxford University in 1988. The transition was managed by Oxford University's Isis Innovation Technology Transfer

¹² US government (NIST's) Economic Assessment Office status report evaluating the performance of their investment in Large Scale Biology Corporation (formerly Large Scale Proteomics Corporation), available at <http://statusreports.atp.nist.gov/reports/94-01-0284PDF.pdf>

Company¹³ and was directed by Raymond Dwek, who later became Head of Biochemistry at Oxford University (2000-06), and is now president of the Institute of Biology, UK.

Oxford Oligosaccharides was a start-up that aimed to identify and analyse a specific type of protein: glycoproteins those modified with carbohydrate molecules¹⁴. They were believed to play a role in important molecular interactions impacting reproductive biology, disease aetiology and regulation of biochemical processes in the body. Initial funds for glycobiology research (£1 million per annum) were provided by Monsanto, then later the VCs, Advent Capital and Euro Ventures (Dwek, 2008). The university, Monsanto (and Searle¹⁵, which Monsanto had then acquired) plus the scientists and staff were the initial shareholders. A board of directors was recruited and head-hunters found the CEO - Raj Parekh, an Oxford postdoc. The company became Oxford Glycosystems in October 1988 and went on to develop and market products for glycobiology, such as GlycoPrep 1000, which in 1992-1994 was "*purchased by nearly every major pharmaceutical company throughout the world*" (Dwek, 2008). However, they did not enter the MS 'proteomics market'¹⁶ proper (as is the subject of this thesis) until they released their first product for high throughput proteomics in 1996.

¹³ Isis Innovation Limited is the University of Oxford's wholly owned technology transfer company. Isis was established in 1988 and manages the University's IP portfolio, working with University researchers to identify, protect and market technologies through licensing, spin-out companies, consulting and material sales.

¹⁴ (Oligo)saccharide is another name for carbohydrate, hence the company name Oxford Oligosaccharides, which developed products related research into proteins modified with carbohydrate molecules.

¹⁵ A company in the life sciences industry, specifically pharmaceuticals, agriculture and animal health. It is now part of Pfizer, the pharmaceutical company.

¹⁶ The 'proteomics market' refers to businesses providing services or products for large scale proteomics, and specifically proteomics that is based on automated mass spectrometry workflows.

Another firm to embrace proteomics, Nonlinear Dynamics, started up in the UK in 1989 in Newcastle-Upon-Tyne, and was to become one of the key players in provision of software for 2D gel proteomics, and is the subject of one of the company case studies to follow.

2.3.2 High-throughput analysis became possible by applying mass spectrometry to proteins

A critical turning point initiating the shift from expensive low-throughput protein identification to affordable high-throughput protein sequencing was the application of mass spectrometry (MS) to proteins (Barinaga, 1989, Hanash *et al.*, 1991, Hillenkamp *et al.*, 1991) (Table 6).

Table 6 Timeline showing key events in the development of high-throughput proteomics and proteomic bioinformatics and funding mechanisms.

	1980s	Early 1990s	Late 1990s	2000s	2008-9 and onwards
Technology developments	2D-PAGE and Edman degradation 'List-based' biology arrives Glycoprotein research begins	2D gel reproducibility improved, quantification using spot size MS for peptide sequencing arrives First search engines arrive on internet Term 'proteomics' coined	Electrospray ionization (ESI) arrives – fully automated workflows now possible Protein biomarker discovery efforts using MS technology Nature warns against 'mindless' high-throughput studies	New methods for quantification of proteins arrive (eg. ICAT)	Validation, statistical approaches to increase quality of identifications derived from automated workflows Standardisation of data and reporting Journals push for data sharing/ dissemination
Funding/ financial support	US government funding (eg. Andersons) UK public funding (eg. Nonlinear Dynamics) Biotech funds (eg. Monsanto/Searle fund Oxford Oligosaccharides)	US government funding Public funding (UK, USA) of research through universities	Protein IP land- grab Private investment (eg. Oxford Glycosciences float, Celera stock offering, Geneprot Darier Hentsch) LSBC acquire Anderson's technology Public funds (eg. Imperial Cancer Research Fund and Mascot)	←—————→ phase Corporate venturing (eg. pharma companies take on proteomics for development of workflows in house) Private investment and deal-making (eg. Confirmant: Oxford Glycosciences and Marconi, Myriad Proteomics) Oxford Glycosciences sold for £102 m Private and public funding (eg. Oxford Genome Sciences South East growth fund)	Public funding of research

In the 1950s, scientists were trying to measure proteins by MS: Klaus Biemann at MIT in, for example, and later Howard Morris¹⁷ at Imperial College. However, the problem of getting proteins to vaporise into the gas phase for entry into the MS instrument was to elude them and the introduction of Fast Atom Bombardment by Michael ‘Mickey’ Barber in 1981 was the major breakthrough allowing native peptides to be ionised, without requiring chemical alteration. However, MALDI¹⁸, the ionisation technique developed between 1987-1991 (Hillenkamp and Karas, 1990) delivered what is now called high-throughput proteomics, because large biomolecules (such as proteins) could be consistently ionised and enter the MS instrument. MALDI-TOF (time of flight) MS arrived in several labs as a bench-top instrument for the first time in the early 1990s. And in 1993, key ideas for applying it to proteins were developed. Indeed, the discovery of the MALDI method earned a quarter of the Nobel Prize in Chemistry (in 2002) to Koichi Tanaka for demonstrating that it could be used to ionise protein.

With new data becoming available from both gene sequencing efforts and new MALDI-derived protein spectra, proteomic bioinformatics was fast approaching. The first ever MS search engine, “Fragfit” (Henzel *et al.*, 1993) performed an *in silico* digest of protein sequences, derived from translation of gene sequences, and compared the resulting peptide masses to MS spectra to identify proteins. Development of this program planted the seed for proteomic bioinformatics.

¹⁷ He claims the first complete sequence of a protein by mass spectrometry (early 1970s)

¹⁸ Matrix Assisted Laser Desorption Ionisation – a method to ionise and vaporise proteins for visibility in MS.

Suddenly freely available software accessible via the internet arrived. Indeed, 1993 was described as a 'vintage year'¹⁹ for peptide mass fingerprinting (PMF) programs, which improved on FragFit's original idea (Mann *et al.*, 1993, James *et al.*, 1993, Pappin *et al.*, 1993, Yates *et al.*, 1993).

As a consequence of the surge in PMF tools, development of new and improved search algorithms began, spearheaded independently by Matthias Mann (Heidelberg), and John Yates III and Ruedi Aebersold (University of Washington, Seattle). These two groups pioneered the first tandem MS search engines, Peptidesearch (Shevchenko *et al.*, 1996) and Sequest (Eng *et al.*, 1994), respectively. These new tools were the first in a new class of proteomic search engines that exploited fragment level information to make peptide identifications from which whole protein sequences could be inferred. Another level of complexity was now being applied to the search; like a Google-type search engine, instead of searching just for webpage titles (peptide sequences), it could now include detailed webpage content (fragment masses) in the search. The knock-on effect was automation of protein identification by combining search engines with high resolution MS instruments (also now becoming available). High-throughput protein identification by MS would soon be affordable to many. No biological insights of great note had been shown, but the technology was now viable enough to publish papers and present results. Scientists were very interested, such as (Kahn, 1995, Jungblut and Wittmann-Lieboldb, 1995), and industry started to notice.

¹⁹ Quoted from John Cottrell, Matrix Science, taken from a tutorial session delivered at the ASMS (American Society of Mass spectrometry) conference, San Antonio, Texas on June 5-9, 2005

2.3.3 Privately-funded biotechs competed for control during the land-grab

In mid-1994, the term 'proteome' was finally coined by an Australian postdoc (Marc Wilkins). Armed with a name, proteomics was ready to enter the main-stream. The land-grab for intellectual property was ready to begin. New biotech companies were established, each determined to be the first to identify proteins and their role in disease, in order to patent them and earn royalties from the pharmaceutical industry. Economics was driving the development of the science now. But it was highly questionable whether biotech entrepreneurs were capable of moving immature proteomics technology from experimental labs to the wider market and be able to deliver knowledge to earn rents (van der Sijde *et al.*, 2003).

One of the first to cash in on the hype was Oxford Glycosystems, who in 1996 achieved over \$60m in private financing (Editorial, 1996) with Kirk Raab (formerly of Genentech²⁰) as CEO. They launched the 'ProteoGraph' product for genome-scale proteomics (Editorial, 1997), and in 1997 formed the 'proteome partnership' with Oxford University, changing their name to Oxford Glycosciences around this time. In April 1998 the company floated on the London Stock Exchange raising £30.8m (market cap: £103m) (Editorial, 1998). This was achieved on their previous success in

²⁰ Now part of Roche, Genentech were the biggest biotech company in the world in the 1990s-2000s. Set up by pioneers of genetic-engineering in 1976.

designing a drug for Gaucher's disease²¹ using glycobiology knowhow. With the funds, Oxford Glycosciences invested heavily back into Oxford University providing a grant of £1.5m to the Glycobiology Institute to set up a proteomics facility in the Biochemistry Department. They then continued in earnest with high-throughput proteomics, not glycobiology.

The imminent arrival of electrospray ionisation (ESI) MS in 1998 further fuelled investment in proteomics. ESI allowed direct coupling of protein separation to MS instrumentation, such that when the instrument was linked to a computer with the pre-processing programs (such as peak pickers) and proteomic search engine(s), the whole workflow from sample to identification could be fully automated. As the Andersons had approached 2D gels in 1980s, now academics and companies in the late-1990s/2000s joined the race to perform huge scale MS-based studies to characterise entire proteomes before their rivals - in genome sequencing style. Having patented their new technologies, in 1999, Oxford Glycosciences, scaled up their operation building a proteomics data 'factory' in Milton Park. Proteomics facilities for biomarker discovery were now emerging in many universities and spin-outs; the pharmaceutical industry had yet to get involved.

In 1997, Geneva Bioinformatics (GeneBio) was formed by Ron Appel, Amos Bairoch and Dennis Hochstrasser, and became the commercial entity representing the Swiss

²¹ A type of lysosomal storage disease, a rare inherited metabolic disorder that results from defects in lysosomal function. Lysosomes are organelles (sub-components) of cells in the body, which digest old organelles, food particles and engulf viruses or bacteria. Symptoms may include enlarged spleen and liver, liver malfunction, skeletal disorders and bone lesions that may be painful, as well as severe neurologic complications.

Institute of Bioinformatics (SIB), providing premium versions of SIB's otherwise free data resources. The company's mission was to provide high quality proteomics databases, software tools and services through in-house development and partnerships with universities, biotechs and pharma companies. GeneBio is one of the company case studies described later.

In London in March 1998 John Cottrell and David Creasy, both having left MS technology firm Finnigan (now part of Thermo Scientific), formed the first ever company based exclusively on a proteomic bioinformatics search engine product, called Mascot (Cottrell, 2003). The product was based on the original MOWSE program (Pappin *et al.*, 1993) developed at the Imperial Cancer Research Fund. A later paper carefully omitted important details of the Mascot product algorithm (Perkins *et al.*, 1999). This company is an interesting case, because it has continued to grow in spite of the proteomics booms and busts around it. It is one of the companies described in detail later.

As commercial labs accelerated their efforts to catalogue all observable proteins using their new pipelines the editor of Nature cautioned that this may not be the most biologically meaningful approach (Editorial, 1999): "*should funding agencies be pouring money into some global [human proteome project] strategy at this point?*" Nature was advocating the small-scale approach, where studies would lead to conceptual understanding of biology, rather than long lists of protein IDs. Quality is better than

quantity. Commercial labs did not take heed of this warning and huge investments were made to achieve the absolute opposite.

Celera, for example, the company responsible for sequencing a large chunk of the human genome, announced their Human Proteome Project in 2000. They intended to identify the properties and functions of every human protein (Butler, 2000) and make revenues through lucrative patents. Craig Venter, the company president, was quoted in Science at the time: *"We're going to have the biggest facility and the biggest database...we'll be working through every tissue, organ, cell"* (Service, 2000). Celera managed to obtain \$944 million in a stock offering (Washtech@WashingtonPost). Venter could achieve these colossal amounts on the back of his previous success in sequencing the genome. The proteome, however, is not the genome; it is something with more dimensions and is much harder to pin down. Venter approached GeneBio co-founder, Hochstrasser, with an attempt to join forces with SIB. No partnership was forged since Hochstrasser's view on open data access did not fit with Venter's vision. Celera is re-visited later.

In 2000 GeneProt (Geneva Proteomics) was established by the three co-founders of GeneBio. It planned to set up a 'proteomics factory' to compete with rival Celera. Huge investment ensued: in April 2000 the seed round of financing²² raised \$4.6 million and three months later a further \$40 million was raised through six

²² Funded through Switzerland's Darier Hentsch Life Science Fund

additional European funds (PressRelease, 2000b). Soon after, they spent \$70 million with Compaq for supercomputing equipment. A deal was then signed with the Swiss pharma giant Novartis, where in return for \$43 million equity investment GeneProt would *“analyse the protein profile of three human diseased tissues...and their healthy counterparts”* (PressRelease, 2000c). The idea was to apply the proteomic data generated by GeneProt to identify novel protein-based targets for use in medicine development. The company was fully operational by 2001, and had an additional facility in USA. At the time, it claimed to have the world’s most powerful super-computer and proteomic discovery facility (PressRelease, 2001f). The investors were leading, the science was trailing.

Next Oxford Glycosciences announced (2001) that they were building ‘ProteinAtlas’. It was to be sold on subscription basis by a company called Confirmant - a joint venture between Oxford Glycosciences and Marconi²³- which formed just before Marconi went into liquidation (PressRelease, 2001c). Their short-term aim was *“to become the leading provider of bio-information”* and long-term, to circumvent pharma distribution channels by developing *“online, real-time diagnostics, made available to physicians...on a pay-per-use basis”*. In parallel to this announcement, the Anderson’s Human Protein Index database was completed. The LSBC annual report for 2000 outlined their strategy. Their database contained protein information for all major tissues so was to become: *“the definitive source of information about human proteins”*,

²³ Available at <http://www.prnewswire.com/cgi-bin/stories.pl?ACCT=104&STORY=/www/story/06-15-2001/0001514667&EDATE=>

although their 2001 annual report stated a shift to *“develop therapeutic products using our proprietary technology”*.

It is clear that the promises of proteomics were fuelling great expectations in the markets. But the market got ahead of itself, before the science could actually demonstrate that it could deliver. According to Frost & Sullivan²⁴ in 2001, proteomics was estimated to grow from \$700m in 1999, to \$5.8bn in 2005 (Frost&Sullivan, 2001). Quotes from the then Director of Drug Discovery Technologies, Stefan Unger, suggest the high-throughput nature of proteomics, and the fact that the technology was in its infancy, were the drivers for the huge investments in proteomic biotechs: *“The main difference between the old and new paradigms is in the high-throughput, parallel thinking...There are no clear winners in these early stages of market development for proteomics, which means there is a wealth of opportunity”*. 151 different proteomics biotech companies appeared in F&S’s 2001 report including all areas of the proteomics market: instruments, “wet” technologies & supplies, lab services, and bioinformatics: *“With over 40 commercial funding activities of various types (VC, IPO²⁵, mergers, etc.), this is a very rapid pace for a discipline that was unnamed just six years ago”*.

In 2001 Myriad Proteomics, was formed in collaboration with Oracle and Hitachi and headed by Nobel Prize winner Walter Gilbert (who founded Myriad Genetics).

²⁴ A global a strategic market research consultancy based in San Jose, CA

²⁵ Initial public offering: when a company issues shares to the public for the first time.

The venture was valued at \$185 million and aimed to “*analyse all proteins and their interactions*” (PressRelease, 2001d)²⁶. The business was based on collection of proteomics information in a proprietary database (to be ready by 2004) using proprietary technologies including: ProNET, protein interaction technology (industry-scale yeast two hybrid²⁷) and ProSpec (proprietary MS technology for identifying protein complexes).

Myriad’s formation marked the ignition of the proteomic biotech bomb. High profile doubters of the high-throughput approach to proteomics began to engage with the media. The president of Hybrigenics²⁸ said of Myriad’s mission: “*there’s no way they’ll come close to it!*” (Pollack, 2001). Myriad replied: they intended to look only at “*10-12 cell types*”. The famous Venter quote “*there ain’t no such thing as a proteome*” then appeared in The Wall Street Journal (Hamilton and Regalado, 2001), and in The New York Times “*We don’t think there’s much value in a general survey of proteins*” (Pollack, 2001). This is the key point, because suddenly everyone had realised there was no ‘value’ in high-throughput proteomics. The economics could not drive the growth anymore, because it was not there. Celera ceased to catalogue proteins, switching instead to designing new drugs as a fledgling pharmaceutical enterprise. LSBC also pulled out: in 2002 they reorganised stimulating both Andersons to resign from their life’s work.

²⁶ Available at http://findarticles.com/p/articles/mi_m0EIN/is_2001_April_4/ai_72721585

²⁷ Interaction proteomics – a method to determine protein to protein binding interactions at the molecular level.

²⁸ A Parisian biotechnology and pharmaceutical company

Oxford Glycosciences was bought by Celltech for £102 m in 2003, which was groundbreaking for a biology-based University spin-out. It led the formation of a new proteomics company, Oxford Genome Sciences, now called Oxford Biotherapeutics (one of the case study firms, see later section). This new company acquired the proteomics division of Oxford Glycosciences and used this to start the new enterprise. This was perhaps not ideal timing to start a proteomics company, since the land-grab was losing momentum and reputation of proteomics was diminishing. In fact, as explained in the case study later, OGS's success would depend on how well they could exploit the potential growth for out-sourcing of proteomics, as the pharma companies began to downsize in-house research in this area in the mid-2000s.

In academia, steps were being taken during the late 1990s to improve proteomics techniques. One of the major drawbacks of proteomic MS was that proteins could be identified, but not quantified (Mann, 1999). To address this, Aebersold and his team invented isotope-coded affinity tagging (ICAT) (Gygi *et al.*, 1999). In 2001, Oxford Glycosciences (before the Celltech buyout) collaborated with these inventors to create state of the art quantitative proteomics facilities (PressRelease, 2001e). Protein interactions were also emerging as a new area of interest. In 2000, proteomic 'interactomics' arrived with yeast-two-hybrid technology being described for the first time (Uetz *et al.*, 2000) (later applied by Myriad Proteomics' ProNET approach, as described earlier). Interestingly, both quantitation and interaction proteomics were moving away from the principle of list-based cataloguing, which was still

going on in industry. Instead, these approaches delivered detailed understanding of sub-sets of proteins, like traditional biochemistry, as Nature has endorsed earlier.

To summarise, it appears that the development of the proteomics market was stimulated by the willingness of investors to believe in MS technology and bioinformatics analysis could deliver. The cost-effectiveness of these new approaches (ability to produce huge datasets in little time) rather than scientific efficacy drove the boom in proteomics biotechs. Indeed, the capacity of new techniques to generate value (as patents on new drug targets, for example) was limited in the extreme. This clearly demonstrates that it was the economics, rather than the true potential of the science that stimulated emergence and growth of this new market.

2.3.4 Big Pharma invested heavily in proteomic bioinformatics infrastructure

In the pharmaceutical industry, heavy investment went into high-throughput proteomics, both during the surge in biotech funding and after biotechs began to fail. During the late 1990s to early 2000s productivity across pharmaceutical R&D organisations was declining (Prasad, 2004, Garnier, 2008). The amount spent on R&D was not reflected in the number of candidates making it through the drug development pipeline (Dimasi *et al.*, 1995); there was an incentive for firms to try out technologies to try to increase the flow. Moreover, in the mid-2000s it was getting harder to convince the regulators that new NCEs (new chemical entities) were safe,

effective and better than the existing alternatives. A new breed of medicines, 'biopharmaceuticals', were now being developed, where large molecules, such as peptides and protein molecules (antibodies, as in Herceptin²⁹, for example) were being designed as therapies, instead of small drug molecules. New avenues, such as these, were being explored because the industry needed innovative approaches; proteomics and its associated bioinformatics activities would form part of this drive.

GlaxoSmithKline (GSK), the sponsor of this EngD (from 2005 to 2009), is taken as a case to illustrate how proteomics was carried out in R&D-based pharmaceutical firms at the time³⁰. Prior to the merger in 2000, both Glaxo Wellcome (Glaxo) and SmithKline Beecham (SKB) were using some tandem MS to identify proteins in samples from humans and model organisms, with both companies having labs in the US and UK. High-throughput proteomics, however, began only after the merger.

Pre-merger Glaxo joined forces with Cellzome for proteomics

With several new proteomics-based biotech companies already trading, Glaxo began proceedings to set up its own independent spin-out company called the Cell Map Incubator (CMI). CMI was based on the biotechnology expertise within Glaxo at the time, and planned to study protein-protein interactions on a large-scale. The CMI company would recruit its scientists from Glaxo, including Walter Blackstock (cell biologist, mass spectrometrists and co-founder of the earlier cell map unit in Glaxo,

²⁹ Trastuzumab is a breast cancer treatment. It interferes with the HER2 protein receptors, which regulate cell growth, survival, adhesion, migration, and differentiation – all processes affected in cancer.

³⁰ Information on amounts invested could not be released. Instead, the information presented is taken from public sources, such as press releases, and from interviews with remaining GSK employees in bioinformatics.

1998), and it would specialise in generating and analysing proteomic MS data. In exchange for their investment, Glaxo would receive proteomics services at a discounted price from CMI in the future.

However, regrettably for CMI, the merger with SKB happened just before it began trading. As a result, several key players pulled out. Some experts remained at GSK, but most moved elsewhere (Malcolm Ward and Helen Byers for example, who joined a new proteomic MS company, ProteomeSciences – listed on the AIM stock exchange in 1995). Blackstock joined Cellzome UK soon after, in 2000.

CMI's business model was based on the expertise, skills and knowledge of its scientists, so after haemorrhaging their most important resource, it was forced to collaborate with an external third party. After deliberation, they chose Cellzome (PressRelease, 2001a). This was a compelling choice, because Cellzome had links with the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. The partnership meant that Cellzome could provide bioinformatics services, and CMI would perform MS in the UK to generate data. This setup appeared to be working until 2006, when Cellzome dropped the MS function, in line with other major downsizing of MS in big pharma. This down-sizing is described now.

Post-merger GSK invested in proteomics workflows and IT infrastructure in the UK and USA

After the spin-out of CMI, a 'home-grown' proteomics facility was established at GSK. Workflows were set up to analyse samples using most types of proteomic analysis; in particular, high-throughput MS to discover biomarkers for a variety of diseases in easily accessible samples, such as serum.

There was a 'buzz' surrounding proteomics at the time. Scientists believed that proteomics would be capable of providing a supply of new biomarkers; so GSK invested heavily in proteomics capability. Indeed around 2002, a major focus was on large scale phospho-proteomics studies, since GSK was aiming to understand how chemical modification of proteins (with phosphate groups) affects biological pathways and cell signalling (Annan, 2002) thus developing new ideas to design drugs to target kinases³¹.

By 2001/2, the six CEDDs³², instituted by CEO Sir Richard Sykes and his successor J-P Garnier were up and running, each investigating a different therapeutic area in GSK's portfolio. Feeding the CEDDs with new candidates were the drug discovery 'research' organisations: discovery research (DR) and genetics research (GR). These

³¹ Kinases are 'druggable' enzymes that catalyse phosphorylation reactions in the body. Druggable refers to their ability to be targeted successfully with medicines. Kinases are phosphotransferases that transfer phosphate groups on other molecules, and are involved in cell signalling.

³² Centres for Excellence in Drug Discovery. These are smaller centres in R&D, each focuses on a specific disease area. They were set up after the merger to improve the efficiency of research in such a large organisation like GSK.

carried out research typical to a “*classical proteomics department*” (Annan, 2002), including comprehensive high-throughput experiments. Organisationally, the bioinformatics group for proteome research was aligned to GR, performing the occasional *ad hoc* project for DR.

Large investment was put into instrumentation (Roland Annan Head of Proteomics: “*We have...one of everything here*” (Annan, 2002)) as well as hardware and software solutions to capture, store, analyse and report huge amounts of data being generated from DR and GR workflows. In the early 2000s, software suites that could pipeline data into a searchable repository were scarce and were either too specific or too generic. GSK built a bespoke, adaptable infrastructure for data analysis, working with a number of third parties, in particular Matrix Science. The result was a system called ‘Proteominer’, directly integrated with the Mascot result files.

Collaborations with academia in proteomics were ongoing, including in 2002 links with Ray Deshaies at the California Institute of Technology (CalTech). No corporate partnerships were in place, however, as was the case for rivals like Novartis, who joined with GeneProt. GSK had greater in-house expertise, for example in genomics, so needed to rely less on third parties (Annan, 2002).

Proteomics was downsized at GSK

The turning point came in 2005. Previously (in 2002), the head of proteomics at GSK (Roland Annan) had stated that “*upper management here thinks that [proteomics] can*

make an important contribution to all aspects of drug discovery...applications for proteomics are still evolving.” (Annan, 2002). Management were changing their view. Annan’s comment on applications of proteomics was telling. It shows that investment has been driven by economic grounds, not based on what science could deliver to GSK. GSK could not find a place for proteomics to add value.

By this time, however, the volume of MS/MS data generated had grown, so too had spending on Mascot licenses and Blade servers. The board had begun to notice that the cost-benefit profile of proteomics was not acceptable. There had been no impact on the drug discovery pipeline; a scenario that was encountered by many labs in industry, as well as their academic counterparts: “*...there are no clear success stories in which discovery proteomics has led to a deployed protein biomarker*” (Rifai *et al.*, 2006).

Technology platforms were prioritised at this stage, and compared with other emerging technologies, such as transcriptomics and high-throughput screening (HTS)³³, proteomics’ impact on the drug discovery was noticeably inadequate, as was the case for GSK’s rivals. Consequently, in winter 2006 all proteomics activity was stopped. The remaining proteomics research activities would be very small or

³³ A method of drug discovery that involves testing hundreds of thousands of compounds against a particular target – so all permutations of the problem are tried to find a match, rather than undergoing rational design. HTS is an automated process involving modern robotics, sophisticated control software, advanced liquid handling and detection methods. The hits generated during HTS can be used as the starting point for a drug discovery effort. Many pharmaceutical companies are screening between 100,000 and 300,000 compounds per screen to produce approximately 100 to 300 hits. Usually only 1 or 2 of these hits become lead compounds for further development. Occasionally, screens of over 1 million compounds are required to generate a sufficient number of lead compounds. HTS can also be used in safety studies, to screen for compounds with undesirable activity.

outsourced. This inevitably prompted disposal of MS instruments and an exodus of MS scientists, several moving to biotechs and back to academia³⁴.

GSK has a new CEO and proteomics does not feature in his plan

Since 2006, proteomics and related research has remained absent from GSK. Current CEO, Mr. Andrew Witty, consulted the firm's shareholders after taking his post in May 2008. Many were discontented with years of underperformance, thus, demanded "*more growth with less risk*" (PressRelease, 2008c). This suggests that the consumer goods part of GSK's business will grow, not the more risky R&D-base in which proteomics and bioinformatics sit (Russell, 2008). Interestingly, Witty's background, unlike previous GSK CEOs, is in economics not science or medicine.

Witty's role as CEO is to mediate between innovation in R&D and the overall business strategy of the firm. As such, he has the central role in the ultimate outcome for GSK. To bridge the market and innovation in R&D, he plans to outsource R&D activities to contract research organisations (CROs); implying that if proteomics were to prove valuable in the future, it would be bought in not grown. He is simplifying the structure of R&D and intends to encourage competitiveness, as in small biotechs, by setting up teams that must compete for up to \$1 billion in annual funds from a 'Drug Discovery Investment Board'. This board will include VCs and an external (biotech) firm's senior executive. The idea of this role-play is to

³⁴ For example, Arthur Moseley who is now the Director of Proteomics for the Institute for Genome Sciences & Policy at Duke University's School of Medicine.

simulate the pitching process of a university spin-out, to ensure that R&D's activities can only proceed if the proposed strategy fits with the current, highly competitive marketplace. This may help, although if the investments made previously by biotech investors in this space are exemplary of the decisions made by the market – then these can also be amiss, presenting Witty with a no win scenario.

2.4 Current status of the proteomic bioinformatics market

In this section the most topical areas of funded academic research in proteomic bioinformatics are described and following this, the current state of the proteomic bioinformatics industry is presented with case studies for firms operating in the UK, continental Europe and North America. In this section, the author demonstrates how the research work for this EngD thesis is relevant and timely, given the status of the field.

2.4.1 Proteomic bioinformatics research is publicly funded again

Proteomic bioinformatics research has continued to receive small amounts of public funding in the UK. Funding awarded between 1998 and 2008 for England is shown in Figure 18 (details of each grant are available in Appendix II).

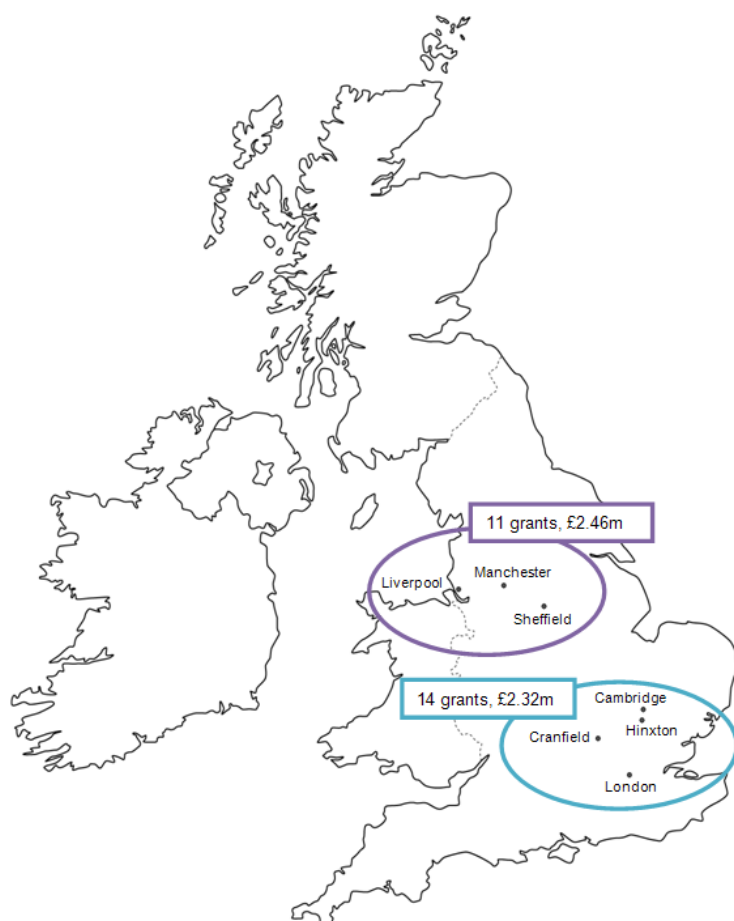


Figure 18 Distribution of BBSRC funds in England for bioinformatics research for proteomic MS between 1999 and 2008³⁵. Only grants for software and computational hardware for proteomics were considered. See Appendix II for information on each grant. From this sample, the Wellcome Trust appears to be the most engaged in funding bioinformatics projects for proteomics, such as the PRIDE (Proteomics IDentifications database) grant (PressRelease, 2005d). The figure illustrates a sample only, other possible funding sources are available, but not reported here due to data access issues.

To obtain a snapshot of public spending globally is more challenging, since there is no easy way to search awarded grants in specific areas, such as via the NIH (National Institute of Health, USA). However, activity in proteomics can be traced indirectly using a major public proteomics repository, The Global Proteome Machine Database (GPMDB), because it provides detailed statistics on hits by country (an example is shown in the Appendix II). It shows the global proteomics ‘hot spots’ are

³⁵ Information taken from BBSRC Oasis database of awarded grants. Only research grants are considered, not teaching or fellowships. The database can be found at <http://www.bbsrc.ac.uk/science/grants/index.html>

Seattle and the South East UK. Seattle is home to the Institute of Systems Biology, and the UK has the European Bioinformatics Institute – major hubs for data in proteomics.

Ideally, large-scale proteomics would now be a commercially viable activity earning rents through with private funding. Instead, it has come full circle. The early innovators (Andersons, Mann and Aebersold) were publicly funded and after the tsunami of biotech and pharma investment, it has now dried up again leaving the field in a funding situation as per the 1990s. This suggests that the economy is still waiting for proteomics to deliver on its potential. Private funding and interaction with financial markets will resume if investors believe in the science of proteomics.

2.4.2 Company case studies illustrate the current market for proteomic bioinformatics products and services

Despite general disinvestment, proteomic bioinformatics companies are still trading. To illustrate the current status of the market, key players have been examined. Case studies are used to examine how businesses were funded over their whole life cycle, and to see if there are any patterns which support or contradict the overall industry summary and the conclusion that too much investment was made too early.

Funding routes are varied, and different funding providers have very distinct expectations as regards the return on the investment they provide. VCs, for example, expect quick and large returns on their investment, as this is the nature of their business model. Government agencies and charities, however, do not usually demand returns in the short term, but rather invest with a view to obtaining revenue

opportunities and societal benefits in the longer term. This difference in priorities must be applied when interpreting the case studies in this chapter.

Only firms that generate revenue from MS-based proteomics software or software-related products and services³⁶ are considered. MS-instrument vendors are therefore excluded because they provide proteomics software *as part of* larger product bundles, and it would be virtually impossible to consider the two revenue streams separately. This may cause a bias, since the MS vendors (such as ThermoScientific, Agilent, Waters, Bruker, Shimadzu and others) represent large providers of software products for MS-based proteomics. In some cases, the innovative elements of the software products sold by these vendors originated from university research or R&D in niche bioinformatics companies and the software is incorporated into MS instrument vendor bundles at a later stage: through licensing agreements, for example. Specific examples include the Sequest search engine which was previously freely available (developed by John Yates III et al (Eng *et al.*, 1994)), but is now exclusively available from ThermoScientific. In this way, the two types of company are linked. Examples of niche firms that have been left out include GenoLogics Life Sciences Software (USA) and Accelrys Software Inc. (Canada); which both have software platforms for proteomics, but their products are not exclusively in this area. Bioinformatics Solutions Inc. (Canada)³⁷ are not included, although they have a search engine product for *de novo* sequencing peptides from MS/MS spectra, because too little information was available to include them.

³⁶ Gel-based proteomics software is also included in some instances. This is acceptable since in practice, gel studies are often linked to MS experiments.

³⁷ See www.bioinformaticssolutions.com

CASE STUDY 1: Oxford Biotherapeutics Ltd.



This start-up company was incorporated in November 2003 and began trading 18 months later. The opportunity to start a new proteomics company came about when Oxford GlycoSciences (OGS) and the newly formed company, Confirmant (a joint venture between OGS and Marconi) were acquired by Celltech plc in 2003. At the point of purchase, OGS had three core areas of business: oncology drug development, inherited disease drug development and proteomics services. The latter two were of significant value to Celltech, offering precious IPR for potentially lucrative drugs. Proteomics, as a research service, did not. This meant that Christian Rohlff (OBT's current CEO) and other senior members of the proteomics division at OGS could strategically acquire technology, infrastructure, IP, bioinformatics and data that Celltech did not want (PressRelease, 2003a), including OGAP® (The Oxford Genome Anatomy Project) (Rohlff, 2004). OGAP is a unique database of high quality protein data derived from (formerly) state of the art proteomics facilities at Oxford Glycosciences, one of the largest in the world at that time. This was not a management buy-out; instead a completely new company was started up. Proteomics was going to be their business, but by retaining a familiar-sounding acronym 'OGeS', it meant that the company could maintain continuity with existing clients and exploit Oxford Glycosciences's good reputation. Five years later - in November 2008 - OGeS became Oxford Biotherapeutics (OBT) to better reflect its expertise in proteomics.

Initial funding for the firm was provided by the South East Growth Fund, part of the UK government agency SEEDA³⁸ in July 2004. Investors in this fund included GE Commercial Finance, Barclays, The Royal Bank of Scotland, the European Investment fund, Berkshire Pension Fund and the DTI (Smart Awards Scheme) (PressRelease, 2005b). The company's pitch was to perform biomarker discovery and evaluation using high-throughput proteomics and bioinformatics. Later the fund invested further (March 2005) (PressRelease, 2005b) as part of a larger investment round including the venture capital firm, Oxford Capital Partners (PressRelease, 2005f). This second tranche of financing was for the move to larger custom-built facilities, the largest of its kind in Europe. In February 2007, Catapult Growth Fund invested £1,200,000 (PressRelease, 2007f). This is private equity firm funded by the UK government's Department for Business, Enterprise and Regulatory Reform and local authority pension funds.

OBT is still privately held and based in Abingdon, Oxfordshire, with an additional site in San Jose, California. It has 16 employees. The firm's income comprises approximately 50% government grants and 50% commercial contracts, see Table 7 for the financial performance information. Grants include schemes such as the DTI's Smart Awards and TSB³⁹ competitions, in return for which OBT must match the donated funds 100% - providing documentation and attending quarterly meetings. The remaining revenue stream is bespoke projects for pharmaceutical companies.

³⁸ South East England Development Agency

³⁹ The Technology Strategy Board. Government agency which has a budget for 2008-2011 of £711 million plus aligned funding from the Regional Development Agencies of £180 million and at least £120 million from the Research Councils.

Since the majority of big R&D-based pharmaceutical companies have downsized internal proteomics research activities, OBT is presented with potentially lucrative opportunities to sell contract research services, allowing their clients to avoid expensive internal headcount.

Table 7 OBT's financial performance (source: Companies House, UK)

Year	Turnover (£)	P & L (£)
2006	427,342	-546,484
2007	0	-1,265,168
2008	not available until October 2009	

The contract services offered include protein biomarker discovery, the “*bread and butter*”⁴⁰ of the business (approximately 60% of contract revenue), and proteomic assay development, including design of MRM (40%). Both of these services involve significant expertise in proteomic bioinformatics and data analysis, and they rely heavily on exploitation of OBT's most valuable asset: OGAP. This is OBT's source of competitive advantage, because it provides the value-added elements to the discovery and development services they offer. The database includes clinical and SNP⁴¹ information along with protein expression data on an enormous scale, containing over a million peptide sequences from over 50 different tissues involved in 58 diseases, including 5,000 cancer membrane proteins⁴². This means that when

⁴⁰ Quoted from Martin Barnes, Head of Bioinformatics, OBT

⁴¹ SNPs and haplotypes are the genetic differences between populations or individuals that can affect susceptibility to certain diseases.

⁴² Membrane proteins are often the preferred targets for drugs

biomarkers are to be discovered many possible clinical implications can be accounted for, allowing evidence-based multi-marker assessment – an attractive proposition to big pharma.

OBT have partnered with Medarex, a specialist in antibody technology. OBT license Medarex's proprietary transgenic mouse⁴³ technology to generate antibody therapeutics against cancer proteins that OBT identify using OGAP. OBT retain worldwide rights to the antibodies generated, and Medarex has the right to receive license fees, milestone payments and royalties on commercial sales of any products that may result from the agreement: all on a 50:50 cost and profit share basis (PressRelease, 2007g). Medarex would have liked to purchase OGAP outright, but greater value could be leveraged by OBT by retaining it.

In 2006, OBT partnered with BioSite (San Diego), a company specialising in commercialisation of protein-based medical diagnosis (PressRelease, 2006a). This strategic partnership improves OBT's position in personalised medicine. There is also a three-way agreement with Biosite and Medarex, where Medarex provide access for Biosite to transgenic technology, and Biosite carry out early stage antibody generation on behalf of OBT for OBT's programs. Amgen also collaborate with OBT to develop and commercialise antibody therapies (PressRelease, 2007a), where OBT's

⁴³ In this case, transgenic mice are genetically modified mice, which are used to generate antibodies that are suitable for human therapeutic use, for example to treat or diagnose cancer. The mice produce 'humanised' antibodies, which are less likely to cause an adverse immune system reaction in humans

role is to provide novel druggable⁴⁴ protein targets to which fully human antibodies can be raised by Amgen's proprietary Xenomouse® (transgenic mouse) technology.

Going forward, OBT wants to commercialise components of OGAP via vendors of MS instruments and associated software packages. MS vendors offer very attractive routes to consumers and could greatly expand OBT's market by exploiting their huge network of existing research laboratory clients. At present, however, OBT is *'too resource-constrained'* to invest in product development to prepare OGAP for this kind of proposition. Of course, OBT's market position is based on OGAP, so they will need to be very careful if they want to share even parts of it with such powerful organisations.

⁴⁴ Implies that the protein can be targeted by a chemical compound (medicine). This type of protein is most suitable when designing a new drug.

CASE STUDY 2: Matrix Science Ltd.



Matrix Science Ltd. was set up by John Cottrell and David Creasy, experts in scientific software / analytical hardware for protein MS, who had prior to their venture worked for Finnigan, now part of Thermo Scientific - a major MS vendor. The company was launched initially in collaboration with the Imperial Cancer Research Fund (ICRF), with the original idea for the product, the molecular weight search (MOWSE) algorithm (Pappin *et al.*, 1993) from Darryl Pappin, head of the Protein Sequencing Laboratory at ICRF. Cancer Research technology, the technology transfer subsidiary of this fund, licensed the rights to MOWSE to Matrix Science. To further develop the program into a viable product, Matrix Science partnered with BioVision (of Hannover, Germany), a peptidomics⁴⁵ company. BioVision had lab-based research expertise, which could be combined successfully with Matrix Science's bioinformatics knowhow. The company was officially incorporated on 24th March 1998, making it a very early entrant into the proteomic bioinformatics market.

The company started with shareholders' funds of less than £40k and is still privately held, with the major partners also being active in running the business. As Cottrell stated in 2003 *"Starting a software company is not as expensive as starting a hardware company, so we did not have to get outside investment, which has given us a certain amount of freedom"* (Cottrell, 2003).

⁴⁵ Peptidomics is essentially the same as proteomics, the difference is that only the peptide sequences are characterised; you do not go the extra step to identify the proteins that gave rise to the peptides.

The company began distributing the Mascot search engine product a year after the company was started. Despite having only seven employees (six in UK, one in USA), MatrixScience has a healthy P&L⁴⁶ account, reporting +£1.67m in March 2007 and +£1.89m in 2008.

Mascot has been around the longest of all the commercial search engines, so many labs and individual researchers are loyal to the product. Commercial data capture and analysis pipelines have incorporated Mascot, via licensing agreements, so it is difficult for customers to switch products easily. Also the target consumer, the research biologist, generally is attracted to a well supported 'black box' analysis platforms, which are easy to use: Mascot fills this niche.

Revenue streams include the sale of licensing agreements to commercial partners for integration of Mascot into their systems, for example with NonLinear Dynamics, IBM, LabVantage's Sapphire, National Institute of Health (NIH), Proxeon and Thermo Scientific. The sale of licenses for their software platforms directly to clients is another major stream. Their products are all related to Mascot and perform proteomic MS data analysis in some form (Table 8).

⁴⁶ Taken from abbreviated accounts at Companies House, UK. A cash flow statement is not included, so no turnover information is available.

Table 8 A summary of Matrix Science’s products

Product	Description	Approximate retail price
Mascot Server	Hardware and software for protein identification	Entry level £4,250⁴⁷ up to £21,800 for eight processors
Mascot Distiller Workstation	Analyses data from multiple vendors includes novel algorithms for peak detection and quantitation	£1,500 plus £750 for each toolbox (e.g. Daemon or quantitation) and £6,000 for Distiller Developer
Mascot Integra	Scalable solution for managing and automating proteomics research, based on Proteominer, which was developed for GSK in the early 2000s	Entry level £20,000
Mascot Cluster	“Turn-key solution” for high throughput protein identification, exploiting parallel computing	Price depends on specification

In addition to the revenue from these products there are also rents earned in the form of support contracts (30% of license fee per year). A final minor revenue stream is the sale of Mascot training courses, which are hosted in various cities across the world and in-house on request. Arguably most proteomics labs in industry and academia, at least in Europe, are Matrix Science customers. In addition to direct sales and marketing, MS instrument vendors perform marketing on their behalf, Bruker Daltronics, for example.

Since 2004, the company has been expanding its Japanese customer base (PressRelease, 2004b) with the establishment of a sister company, Matrix Science KK, which was assisted by The Japan External Trade Organisation (JETRO), a government-funded organisation that promotes inward investment.

⁴⁷ Prices were by quotation and exclude VAT

CASE STUDY 3: Nonlinear Dynamics Ltd.



Incorporated in Newcastle-Upon-Tyne, UK, on 9th October, 1989, Nonlinear Dynamics is a privately owned company specialising in analysis and data-mining tools for 1D and 2D electrophoresis gels and for MS. Nonlinear is a family-backed business with former 'directors' including William, Sheila and Alfred Dracup and the current CEO is Will Dracup.

Currently with 29 employees and an additional office in North Carolina, USA, Nonlinear is one of the larger companies in this market. The firm was financed initially through bank loans (e.g. £15,000 in 1993), followed by equity investment by British Coal in 1995. Funds from the Smart Awards scheme⁴⁸ were granted, but the majority of investment was provided by Northern Enterprise⁴⁹. NEL Capital Fund Managers also have Nonlinear in their investment portfolio⁵⁰. In 1995, the first distribution contract was signed and in early 2000 Nonlinear launched their most successful product, Progenesis. The end user price for this software increased from £4.5k to £80k and the company headcount grew suddenly as a result, growing from just 25 to 100. However, revenue streams dried up in more recent years and redundancies ensued (Figure 19).

⁴⁸ Smart Awards were given to individuals and small and medium-sized companies (<250 headcount) by the Department of Trade and Industry. The last one was awarded in 2003 NEWSLINK (2003) 'Smart' companies get their rewards *Newcastle University's Newslink*.

⁴⁹ From Will Dracup's blog at <http://www.ifwecanyoucan.co.uk/Entrepreneurs/Will-Dracup/my-story>

⁵⁰ NEL is a VC company investing in high growth businesses in North East England <http://www.nel.co.uk>

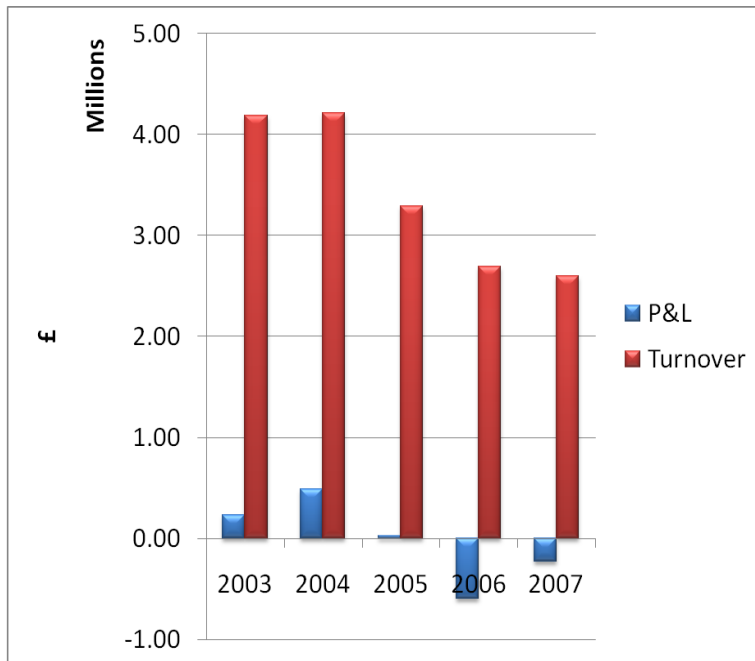


Figure 19 Nonlinear Dynamics financial performance⁵¹

Nonlinear’s customers are universities and academic institutes and also Novartis and BioMerieux. They are also a distribution partner for GeneBio, distributing Phenyx globally (PressRelease, 2008e). For a summary of their products see Table 9.

⁵¹ As reported in the Fame database

Table 9 A summary of Nonlinear Dynamics's products

Product	Description	Approximate retail price (in \$)
Progenesis	software for 2D gel, DIGE, LC-MS and biomarker screening	22,000
Progenesis SameSpots	2D and DIGE analysis platform	24,000
Progenesis Stats	multivariate statistical analysis tool for 2D and LC-MS	Unknown
Progenesis PG600	biomarker discovery using MALDI TOF MS analysis	Unknown
TL100	quantitation and calibration of 1D gels	999
TL120	analysis tools for quantitation, calibration and band pattern matching	3,000

Nonlinear has two wholly owned subsidiary companies: Nonlinear EBT Limited and Phoretix International Limited, and has distribution partners in Korea (Chayon Laboratories Inc.) and Japan (SCRUM Inc) (PressRelease, 2007e). In August 2004, Nonlinear sought collaboration with Matrix Science Ltd. Under this agreement, Nonlinear's protein informatics system was integrated with the Mascot Server product. They also have agreements for distribution with Perkin Elmer, who have proteomic gel imaging products (PressRelease, 2005e).

CASE STUDY 4: GeneBio S.A.



Geneva Bioinformatics (GeneBio) was founded in 1997 by three professors from the University of Geneva: Ron Appel (Department of Computer Sciences and Executive Director of the Swiss Institute of Bioinformatics⁵² (SIB)), Amos Bairoch (Department of Structural Biology/ Bioinformatics and head of the SWISS-PROT⁵³ database group at SIB) and Denis Hochstrasser (director of the Clinical Pathology and Vice Dean of the Faculty of Medicine).

The current CEO, Nasri Nahas, took the role in 2001 having acquired experience in the biotech industry at Genset SA, which specialised in genomics⁵⁴ and was the second largest biotechnology company in Europe in 1999, and also ValiGen SA, a EuroAmerican functional genomics⁵⁵ company. Former director, Prof. Robin Offord (1998-2000) came from GeneProt. GeneBio was funded predominantly by Index Ventures, a pan-European VC firm focused on the life science and information technology markets (PressRelease, 2001b).

⁵² SIB is an academic not-for-profit foundation established on 30th March, 1998. It coordinates research and education in bioinformatics and provides bioinformatics services for various areas of biology to international research communities via the internet.

⁵³ SWISSPROT is a protein sequence database hosted at SIB, which was first created in 1986 by Amos Bairoch during his PhD. It is a manually-curated database which means it provides a high level of annotation (such as the description of the function of a protein, its structure, chemical modifications and variants). It has little redundancy, making it more compact than other protein databases available in the public domain. It also has links to relevant external data resources.

⁵⁴ Genomics is the study of genomes, where a genome is the entire DNA sequence of an organism.

⁵⁵ Functional genomics aims to understand the function of the genes that make up the genome, so includes research into on the dynamic processes such as gene transcription and translation – which leads to protein production in cells.

In 1998, GeneBio became the exclusive commercial representative for the Swiss Institute for Bioinformatics (SIB), the developer of key proteomic tools and protein databases like SWISS-PROT, PROSITE⁵⁶ and SWISS-2DPAGE⁵⁷. In May 2004, GeneBio expanded into Japan (PressRelease, 2004a). In 2005, another company was founded called Current BioData, Ltd (CEO Ian Tarr) as a joint venture between GeneBio and the Current Science Group (London). The company would focus on the further development, promotion, and distribution of the ProXenter product. Current BioData set up a research site in Wales in 2008 (PressRelease, 2008a).

Financial performance information is unavailable for this firm. The company's products (Table 10) are priced on an individual basis.

Table 10 A summary of GeneBio's products

Product	Description
Phenyx	MS data analysis platform released in 2004. It is sold as PhenyxServer, with a CPU-license (set price per CPU) and as PhenyxOnline with an annual subscription the price of which depends on the “user profile” (i.e. number of monthly submissions, size of peaklist files,etc.)
SmileMS	Metabolomic MS data analysis
Melanie	2D gel analysis which is sold as a user-license (price per user number)
Aldente	For PMF sold as a PC-license (price per PC)
MSight	For graphical exploration of huge MS datasets
e-proxemis	Bioinformatics learning portal (launched 2005)
Premium versions of the SIB Databases	User license (with a set price per user number)
ProXenter	Web-based information portal (released as a joint venture with Current Science Group)

⁵⁶ PROSITE is a manually-curated database of protein families and domains hosted at SIB. It was set up in 1988 by Amos Bairoch.

⁵⁷ SWISS-2DPAGE is a database of annotated 2D and 1D PAGE gels hosted at SIB, set up 1993.

GeneBio's key customers are academic institutions, such as the Biozentrum at the University of Basel (PressRelease, 2007b) and the proteomics group at Utrecht University (PressRelease, 2009). They are well-connected with partnerships including Nonlinear Dynamics, GE Healthcare, Bruker Daltonics, Genedata, Amersham (PressRelease, 2003b); Wiley-VCH (PressRelease, 2005g); Sage-N Research (PressRelease, 2006e); Insilicos (PressRelease, 2006b); Genologics (PressRelease, 2006c); Proteome software (PressRelease, 2007c); Institute of Systems Biology, Seattle (PressRelease, 2007d); Protagen AG (PressRelease, 2008f); and Proxeon (PressRelease, 2008b). GeneBio has international partnerships including KOOPrime Pte (Singapore) (Summary, 2005); BIGG (The Bioinformatics Institute for Global Good (BiGG), a research institute in Tokyo, Japan) and Hitachi Software Engineering Ltd. (PressRelease, 2005c); Proteomic Solutions (France) (PressRelease, 2005a); and Proteome Systems (Sydney, Australia) (PressRelease, 2000a).

CASE STUDY 5: Proteome Software Inc.



Mark Turner and Brian Searle⁵⁸ started Proteome Software Inc. in 2004 in Portland, Oregon, USA. Mark Turner's background is in information technology and previously (1994-1996) established a successful start-up called Noetix based around views for visualising data in commercial Oracle databases. Brian Searle was originally trained in chemistry, but then later moved into proteome informatics research and software programming. The company was initially funded internally and has since been funded entirely on sales.

Before forming their company, Searle and Turner had worked together at Oregon Health and Sciences University under Srinivasa Nagalla, MD⁵⁹. Here they developed proteomics software called OpenSea (Searle *et al.*, 2004, Searle *et al.*, 2005) to interpret peptide *de novo* sequencing data. OpenSea was intended to be Proteome Software's first product but IP issues complicated the spin-out process.

Instead, the first year was spent developing an alternative software product called Scaffold. This software package helps scientists interpret results from proteomics search engines such as Sequest and Mascot in a consistent and reliable manner. To achieve this, they re-implemented and pipelined strategies from the proteomics

⁵⁸ Brian Searle was interviewed by the author on 15th July, 2009, Cambridge, UK. This case study text has also been edited directly by Brian Searle (31st August, 2009).

⁵⁹ Founder and CEO of Diabetomics, a medical diagnostics company. He is a pioneer in application of genomic and proteomic technologies for medical diagnostics.

literature, most notably PeptideProphet (Keller *et al.*, 2002a) and ProteinProphet (Nesvizhskii *et al.*, 2003b) for determining peptide and protein identification probabilities.

Once the product was ready for market, the company hired Mark Pitman as a primary sales lead. Turner took on management and product testing responsibilities and Searle took technical development. Searle admits to using Matrix Science as a model for setting up and maintaining a profitable proteomics software company. Proteome Software has gone on to specialise in developing tools to make complex data analysis easier for lab-based researchers.

In 2009, the company is still privately held and employs just nine people: one manager, three software developers, two sales leads, two customer service managers, and a software tester. The company has a modest list of specialised products that are sold to industry and academic labs worldwide, but predominantly in the USA and Europe (Table 11).

Table 11 A summary of Proteome Software's products

Product	Description
Scaffold 2	Interpreter for MS/MS based proteomics that combines and compares multiple samples and database search engines into a single experiment-wide view
Scaffold Q+	Relative quantitation tool for MS/MS based proteomics
MassQC	Online service that stores, analyzes and displays performance metrics for LC-MS/MS based proteomics for quality control

Sales are made directly, or via resellers, such as Mass Solutions Technology in Taiwan, Software4Labs, UK, and Matrix Science KK, Japan. Proteome Software do not publish accounts. The funding structure, however, is known to be based on reinvestment of profits, for example into new product development, rather than requiring external investment - at least for now.

CASE STUDY 6: Sage-N Research Inc.



Sage-N Research's CEO is David Chiang, a member of the Sand Hill Angels⁶⁰, an early-stage venture and mentor capital firm. He is an inventor and engineer, educated at MIT, and before founding Sage-N Research, he was employed by Xilinx, Inc., the world leader in the Field Programmable Gate Array (FPGA) industry⁶¹. Before this, he was a Senior Design Engineer for Altera Corporation, as well as for the Research Laboratory of Fairchild Semiconductor.

The company was incorporated in 2002 and is in San Jose, California, and Shanghai, China (since December 2005). The Silicon Valley connection is unique in the proteomic bioinformatics field, and Sage-N plays this to its strength with its major product being the only integrated data appliance for proteomic MS-based research. It also boasts the biggest names in proteomics on its advisory board, including Zubarev, Aebersold, Gygi and Yates. Sage-N's major revenue stream is the development and sale of platforms for high-throughput proteomics data analysis (Table 12).

⁶⁰ <http://www.sandhillangels.com/>

⁶¹ Information is from the Infotrac Company Profile for Sage-N, dated 5th Dec 2008

Table 12 A summary of Sage-N Research's products

Product	Description
Pattern Match Accelerator	Hardware and software that are as powerful as a computational cluster but are a single machine
Sorcerer 2	Software that supports continuous high throughput proteomic data analysis. It is extensible, accommodating Open Source as well as proprietary analysis algorithms.
Sorcerer XT	Rack mounted server version
Sorcerer Enterprise	Scalable version for large research centres, designed for Linux platforms
Sorcerer's Shield	Subscription program (for continuous updates and zero down time) renewed on a yearly basis for up to five years (\$5500/year) are additional add-on lines

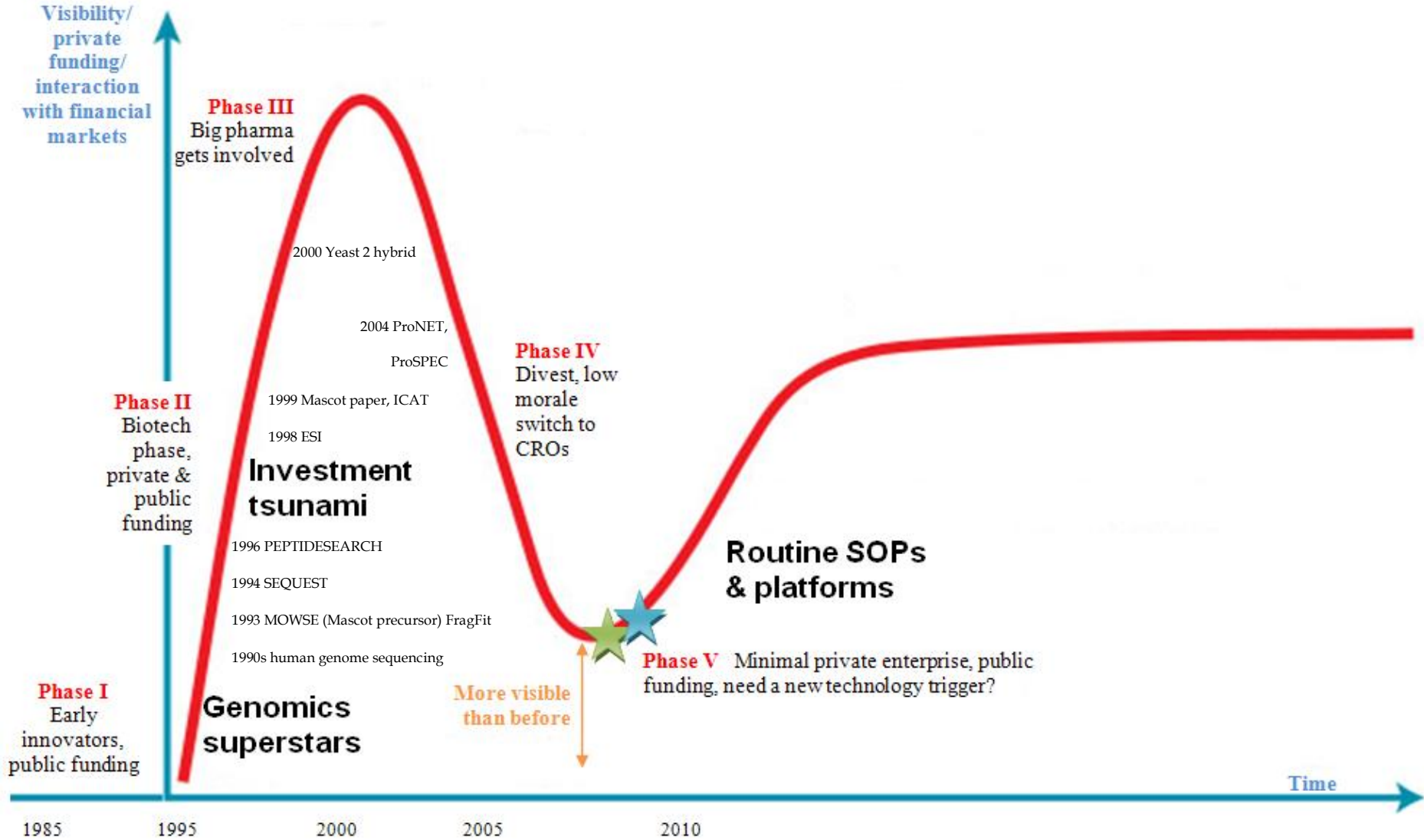
Sage-N have an agreement with Proteome Software Inc. (PressRelease, 2006f) for distributing Scaffold as part of the Sorcerer bundle. In 2007, they were also granted a sublicense from ThermoFisherScientific to sell SEQUEST® search engine (PressRelease, 2007h, PressRelease, 2008g). Other partnerships include VM Ware, IBM, Institute of Systems Biology (PressRelease, 2005h); Rosetta BioSoftware (interoperability between Elucidator® and Sorcerer) (PressRelease, 2006d) and GeneBio (Sage-N can distribute GeneBio's Phenyx platform together with Sorcerer) (PressRelease, 2006e). Sage-N's customers include universities, institutes and private companies in USA, UK, Israel, Canada and Singapore.

2.5 Discussion

Proteomic bioinformatics is a new field of science. The chapter so far has explained how it developed and how it was funded. In the discussion, the business history and case studies are interpreted, and recommendations are made to fragment the proteomics value chain. This will increase efficiencies in R&D spending on proteomics in big pharma and, in turn, will generate further growth potential for niche proteomic bioinformatics companies.

2.5.1 High-throughput proteomics is a typical 'hype cycle' technology

Gartner's technology hype cycle offers a useful model for interpreting the series of events seen in proteomics (Figure 20). The x-axis is time, and the y-axis is 'visibility', which represents private funding and the level of interaction with financial markets in general. The model shows that there is a direct connection between the way science is funded with how it evolves.



1985 protein catalogue using Edman

1987-91 MALDI,

Early 1990s MALDI-TOF

Figure 20 Gartner's technology hype cycle applied to proteomics. The green star marks the current location of high-throughput proteomics technology (MS/MS and associated bioinformatics) in pharma; the blue star represents the location of academia and other publicly-funded research organisations. Process innovations are annotated on the curve by the year(s) they emerged. Gartner is a large, US market research firm for high-tech and IT. (Source: images from Wikimediacommons, Matthias Mann's MaxQuant website and <http://www.sttammany.lib.la.us>)

Phase I: The technology trigger came from publicly-funded research

Prior to proteomics taking off in industry, two developments had begun in publicly-funded organisations: (1) in the late 1980s and 1990s researchers began large-scale cataloguing of proteins using existing low-throughput methods; then (2) new technology (MS and proteomic search engines) was invented which could increase the throughput of cataloguing, meaning proteins could be measured on an industrial scale.

These catalogues were effectively proteomics 'taxonomies', where the aim was to record and measure every human protein. This was an obvious thing to do first: like explorers, or alchemists, endeavours were made to carefully characterise the landscape and give the new field boundaries. An effort to create such a catalogue is perhaps unlikely to create commercial value in the short term, but the public funders appreciated that this was uncharted territory and they were willing to support it. Indeed, funding of fundamental research like this is justified by governments by referring to "*indirect but important long-term benefits to society*", with the division of labour with firms coming later once technical advancements have been made (McKelvey, 2000).

Phase II: Industrial scale proteomics was possible and investors believed the genomics 'superstars' could deliver

The transition from publicly funding to private investment came about because:

1. There was a technology push: industry-scale proteomics was now possible, specifically via MALDI (1993) and ESI (1998).
2. IP was potentially available using the new technology; there was a desire to increase the effort for protein biomarker discovery in pharma and biotechs.
3. Genomics was developing high-throughput science 'superstars', who attracted the attention of investors.

During the land-grab, a boom began for companies to create their own catalogue of proteins and mark their territory in the form of IP. This was fundamental exploration of the science way before commercial viability had been demonstrated.

During the 1990s, the greatest taxonomy yet - the human genome project - was almost complete. The sequencing effort for genes began in 1990, a draft was available of the entire genome in 2000, and by 2003 it was complete. And in contrast to proteomics, high-throughput sequencing of genes created commercial value almost immediately, in the form of low cost microarray products for measuring gene expression using the sequences taken direct from the genome sequence. In 1997, for example, the first miniaturised microarray was created (Lashkari *et al.*, 1997), with commercial versions coming soon after: Affymetrix⁶² in 1994, Agilent⁶³ in 1997.

⁶² Scientists at Affymetrix invented the world's first high-density microarray in 1989 and the company was the first to market with a DNA microarray.

⁶³ Agilent is a Hewlett Packard spin out . In 1997 as HP, it introduced its first microarray product (the GeneArray Scanner) for analysing GeneChip probe arrays from Affymetrix.

As a result of the incumbent success of genomics in the mid to late 1990s, the new breed of genomics biotech 'champions' (as defined in (Markham, 2002)), like Venter, Hochstrasser, Gilbert and Raab - who had been highly visible figures in gene-related enterprise - were powerful, well-connected and respected in both commercial and scientific spheres. They effortlessly attracted the confidence of investors when they set out to launch proteomic biotech companies. However, the land-grab did not succeed. Expectations were artificially high; compared to genomics, the problems being addressed were more complex: "*the proteome is analogue, the genome is digital*" (Moody, 2004). The human proteome, as a concept still is not fully defined, whereas it took less than ten years to have the idea and complete the entire sequencing of the whole genome. The number of genes is 20,488 (Clamp *et al.*, 2007), but there are more than 200,000 proteins including variants and possibly up to ten million if all somatic DNA rearrangements are included (Uhlen and Ponten, 2005).

For evaluating the potential of biotech companies, JP Morgan suggest that three requirements need to be satisfied (Figure 21): people, IP and science (Berry, 2002).

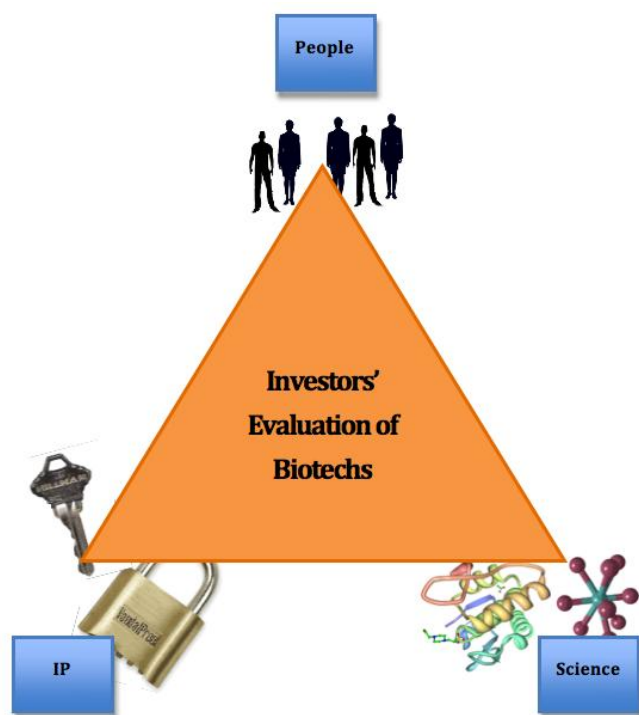


Figure 21 Investors weigh up biotech start-ups using three main criteria: people, IP and science (Source: JP Morgan in (Berry, 2002))

It could be argued that the proteomics biotech of the late 1990s easily met the people criterion that investors were looking for, but the IP rights and the science were notably absent; spelling their downfall. The reputation of the 'genomics superstars' sold the idea of high-throughput proteomics, despite the 'experts' having specialised knowledge derived from a distinct field. One of the challenges facing middle/senior managers is how to determine the economic value of the new technologies and knowledge they have or promise to deliver (McKelvey, 2000). At the time, it was impossible to understand the proteome in enough detail to patent or just apply elements of it to patents or product development. Fishing for marketable biomarkers in the 'spectral soup' was hindered by the technical difficulties, such as

background noise, orders of magnitude differences in protein expression and issues with reproducibility. Sophisticated tools were needed to integrate and interpret proteomics data, and many different groups needed to coordinate to get the job done.

There was not a public policy issue to prevent IP for protein biomarkers. Indeed, despite President Clinton and Prime Minister Blair announcing in 1999 that gene sequence information should be made freely available, thus not patentable (Hencke *et al.*, 1999), by spring 2000, Clinton had clarified the statement: *“if someone discovers something with a specific commercial application, they should get a patent”* (Pilling, 2000). And in any case the genomics market had grown on the back of technical microarray products, not just gene leads *per se*. Protein sequences and information could in theory be patented, as demonstrated by Oxford GlycoSciences who successfully patented 800 proteins associated with human disease in 2000 (Firn, 2000), it was just that there were very few leads to patent, because the technology could not deliver them. In summary, the IP dimension of the triangle (Figure 21) was not to blame. It was the science criterion of the model, there was a problem. Since industry-scale research of proteins became possible, it was automatically deemed worthwhile doing it, but *“technology should not be about higher throughput,...but a means to provide insight...of the biology...under investigation”* (Naylor *et al.*, 2007). The science was not ready to deliver knowledge with economic value.

Phase III: The peak of inflated expectations - proteomics was driven by hype and rivals in corporate pharma

Corporate pharma entered the market for proteomics relatively late. There was increasing pressure to be seen to be investing, since “... *not doing anything is generally not a real option for firms engaged in fast-moving environments...[it] means being left behind and out of business*” (McKelvey, 2000). By investing in high-throughput proteomics, GSK could demonstrate that it was not missing out in the technology potentially delivering value to their rivals. Operations in proteomics were at full speed (at GSK, for example) at the very tip of the wave (Figure 20).

Companies, like GSK, have a history of ‘punting’ on several new technologies in parallel in R&D and then assessing the value added by each after a period of time. For them, proteomics and its associated bioinformatics activities was just another component in a very large ‘machine’ (Garnier, 2008). Unlike the biotechs, it did not spell bust, because there were other elements of the business that could keep the machine running. However, proteomics is an example of how technology alone is unable to cure the inefficiencies in R&D: “*without true understanding of ...new technologies...and the ability to interpret the complex and massive datasets that are produced...how can we expect [technology] ...to cure...all [the pharma industry’s] woes?*” (Naylor et al., 2007).

Phase IV: The trough of disillusionment - proteomics has no interaction with the financial markets

Industrial proteomics in the mid to late 2000s was noticeably absent from the headlines, with no interaction with the financial markets: *"Once the poster child of the biotechnology revolution... seen as the next technological gold rush, proteomics has gone very quiet."* Russ Swan⁶⁴, 22nd July 2008. Given the very steep increase in investment seen in the wave, proteomics may be considered the victim of its own successful marketing. However, *"...much of the data generated by proteomics groups over the past decade is junk."* (Service, 2008a). The plasma proteome project, for example, was *"a big disaster"* (John Yates quoted in (Service, 2008a)) because of the lack of quality control and reproducibility obtainable with MS. Moreover, downsizing cost scientists (and middle management) their jobs, morale and credibility.

As a hit back, the Nonlinear Dynamics CEO, Will Dracup, is leading the Fixing Proteomics Campaign⁶⁵ (2007) to protect the marketplace (including his own company's) and to *'bring together the people in proteomics who want to tackle the growing frustration and unfair perception that proteomics hasn't delivered'*. How 'unfair' the perception is can be debated, since as yet there are still no commercially valuable biomarkers derived exclusively from proteomics (Rifai *et al.*, 2006).

⁶⁴ Editor of the Laboratory Talk Blog (<http://www.laboratorytalk.com>)

⁶⁵ <http://www.fixingproteomics.org/>

Phase V: The slope of enlightenment – a new technology trigger or consolidation of existing techniques?

In spite of the difficulties, proteomics still has the potential to become a big market. Unlike genes, proteins are the physical targets of drugs; they carry out the biological process in all cells in all organisms, so there is great potential for leveraging understanding about them to create commercial value. *“The lure of proteins was undeniable”* (Service, 2008a) and it still is. For now, however, high-throughput proteomics is in the trough, albeit higher on the y-axis than when it started out (orange arrow, Figure 20). Possible options for escape are the arrival of a new technology trigger, or revisiting existing techniques applying learning from mistakes. For the first scenario, the limits of MS and current database searching techniques must be overcome. It may be that investors will need to see totally new technologies that have been proven, to lift the market out of the dip and start a new cycle. Given the poor track record for proteomics, funding will only be made available if scientists can reliably substantiate their claims. Furthermore, given the economic crisis in the financial industry during 2008-9, governments and private companies have contracted budgets further; President Obama, however, appears to be backing fundamental scientific research with \$1.1 billion already promised for research funding to NIH and other bodies (Mundy, 2009). However, there is a noticeable absence of ‘superstar’ scientists to lead the cause this time, so revolutionary new techniques may be even harder to sell.

If, in contrast, high-throughput proteomics is revisited in a concerted effort, then best practice and standards must be set and adhered to. This is the route of escape that the author believes is happening now. For example, since January 2006 the US National Cancer Institute has been running the Clinical Proteomic Technologies for Cancer (CPTAC)⁶⁶ programme: a five year initiative to develop standard operating procedures (SOPs) for high-throughput proteomics (PressRelease, 2008d, Blow, 2008). The program will cost \$104 million; and they have already spent approximately \$35.5 million. Five labs are involved in setting SOPs for unbiased biomarker discovery and biomarker verification with MRM including bioinformatics analysis. A new tool for MRM transition design, for example, was funded by CPTAC (Skyline from the MacCoss lab (Prakash *et al.*, 2009)) (personal communication, MacCoss, June 2009). Furthermore, recent news articles hint that leaders in proteomics research, such as Matthias Mann, Mathis Uhlen and Amos Bairoch, are forming a plan to undertake a new full-scale human proteome project (HPP) (Editorial, 2008a, Service, 2008a), estimated to cost in excess of \$1 billion and take 8-10 years. This may not happen just yet, as European-wide funding is harder to coordinate. Nevertheless, a pilot study is being considered for mapping the proteins expressed by genes on chromosome 21.

Indeed, it seems that the profile of high-throughput proteomics is improving. €12 m over five years (2008-2013) has just been awarded by the EU framework 7⁶⁷ to the

⁶⁶ http://proteomics.cancer.gov/about/CPTC_milestones_508.pdf

⁶⁷ Seventh Framework Programme of the European Community for research, technological development and demonstration activities (http://cordis.europa.eu/home_en.html)

PROSPECTS (PROteomics SPECification in Time and Space) project (Cottingham, 2008), and the Science magazine 2008 'Breakthrough of the year'⁶⁸ runner up was Matthias Mann and co-workers' with work in large scale proteomics. Mann *et al.* can now identify the complete yeast proteome "*in one shot-in just a few days*" using tuned MS (Service, 2008a) and new software, MaxQuant (Cox and Mann, 2008); a feat which in 2003 would have taken months five years earlier. Their strategy is truly like a gene sequencing facility - running samples constantly in a fully automated fashion. The difference this time is that the cataloguing approach has impact on science. For example, they can measure all of the proteins expressed in cells lines and knock-outs, which are used by the pharmaceutical industry to test new medicines. Specific biomarkers are not here yet, but reproducible platforms are emerging.

Phase VI: Plateau of productivity - routine and affordable proteomics technologies are needed

For the technology to plateau, the technology must become routine and affordable. To this end, efforts like CPTAC, and Mann's work brings the field a step closer to an affordable method to map the proteome and exploit it to examine global changes in protein expression for delivering biomarkers. For proteomics to see growth, like genomics, the industry needs a single, affordable technology platform to tell researchers all they would like to know about proteins.

⁶⁸ <http://www.sciencemag.org/cgi/content/full/322/5909/1768>

From the viewpoint of the author, high-throughput proteomics is now beginning to ascend the slope of enlightenment, but has not yet reached an acceptable plateau. The aforementioned efforts work towards the formulation of SOPs and best practice (CPTAC, Mann, *et al.*). Furthermore, the author argues that the tools and resources developed in this EngD add to the progression of this upward slope. For example, by increasing confidence in automated searches (in Chapter 4) higher quality is established in high-throughput analysis workflows; it acts as a way to establish best practice and suitable reporting guidelines for protein identifications derived from pipelines. Indeed, new statistical techniques and methods to increase the quality of identifications are topical right now (Elias and Gygi, 2007) (Nesvizhskii *et al.*, 2007) (Choi and Nesvizhskii, 2008a, Käll *et al.*, 2008, Tabb, 2008).

Furthermore, the computational resources for MRM (developed in Chapter 5 and 6) are important steps towards the future of proteomics, because it is highly likely that MRM will feature in the final toolbox of SOPs for proteomics, and in routine clinical proteomics practices on the plateau. The MRM tool and database developed in this thesis fit with CPTAC's aims to improve the quality of biomarker validation using **targeted** and **reliable**, and **quantitative** approaches - not high-throughput shot-gun-style proteomics seen previously. In fact, the MRM tool published as part of this thesis was released before the CPTAC offering (Skyline), showing the timeliness of the efforts presented by the author.

2.6 Proving the management hypothesis

The hypothesis for this investigation was that the economics was leading the true value creator: science and technology.

The business history of proteomics showed clearly that despite the advent of genome sequencing (leading eventually to the development of the first search engines), scientists like the Andersons, were aiming to catalogue proteins years before a genome was available. So proteomics started out in the absence of genomics. It started, like other sciences, as publicly-funded exploratory science. Only when it became possible to perform large-scale proteomics, through technologies such as bioinformatics (now using genome sequences) and high-throughput LC-MS set-ups, did private investors get involved. When they did, it was frenzy and a new market was rapidly cultivated. It was a **technology push** scenario, where the *improvements in technological ability* to analyse proteins fuelled the investment in proteomics and bioinformatics, rather than a market pull where technical improvements were made *in response to a perceived market demand* for protein analysis.

In contrast, however, pharmaceutical companies had declining productivity in R&D and were looking for ways to fill their drug pipelines with new targets. Proteomics was one of many new technologies offering potential use for biomarker discovery so

it fitted the bill. In contrast then, a **market pull** was driving the investment and growth in proteomics in corporate pharma and indirectly affected biotech growth.

Did the market for proteomic bioinformatics come about because of process innovation, or did it come about because of the excitement and the success of the genome sequencing project? It is true to say that the excitement and the success of the genome sequencing project helped to fuel the growth in commercial proteomics in biotechs and big pharma in the late 1990s- early 2000s, in that this era produced the genomics 'superstars' who caught the eye of investors and hence, socio-economic interactions ensued. However, availability of genome sequences *per se*, had little effect on the growth in proteomics at this time. Since when one search engine was developed to demonstrate the principle of applying genome sequences to proteomics that was it. Search databases did not change significantly after this; instead developments were made in instrument design and software algorithms. It was the *reputations* of individuals from the genome sequencing era that were more responsible for the growth in proteomics, than the improvements in genome sequences themselves.

The author argues that increases in processing power, new algorithms (such as Mascot), and capability of MS as a technique, did have an impact on the attractiveness of proteomics and hence the emergence of proteomics and proteome bioinformatics as a new market in its own right. But, by the boom in the market in the late 1990s, the science could not keep up with the demands of the market.

2.7 Recommendations for investors and funding bodies for proteomic bioinformatics

The hype cycle demonstrated that proteomic bioinformatics is a high-tech industry with close parallels with the IT industry. Indeed, the dot-com bubble (1998–2001) was happening at the same time as the boom in high-throughput proteomics biotechs. By 2001, the IT industry was in crisis, but it picked up quickly afterwards with success stories like Microsoft and Google, and more recently Skype and Facebook. So, for investors looking to invest in the next phase in proteomic bioinformatics, what recommendations can be made? To answer this, the trends in funding mechanisms are analysed, and suggestions made given the author's interpretation of the results of the analysis.

2.7.1 Funding high-tech can be a springboard for growth or a futile cycle

The typical funding stages for high tech companies are shown in Figure 22. High-tech is a high risk, potentially high-growth industry, so VC funding is usually required. Funding is multi-stage and exit routes (of interest in particular to the VCs involved) are: a buyout, for example by a larger corporate firm; flotation on the stock market; or (in the worst case) a cycle where start-ups require further government support to continue development and trading (blue arrow, Figure 22).

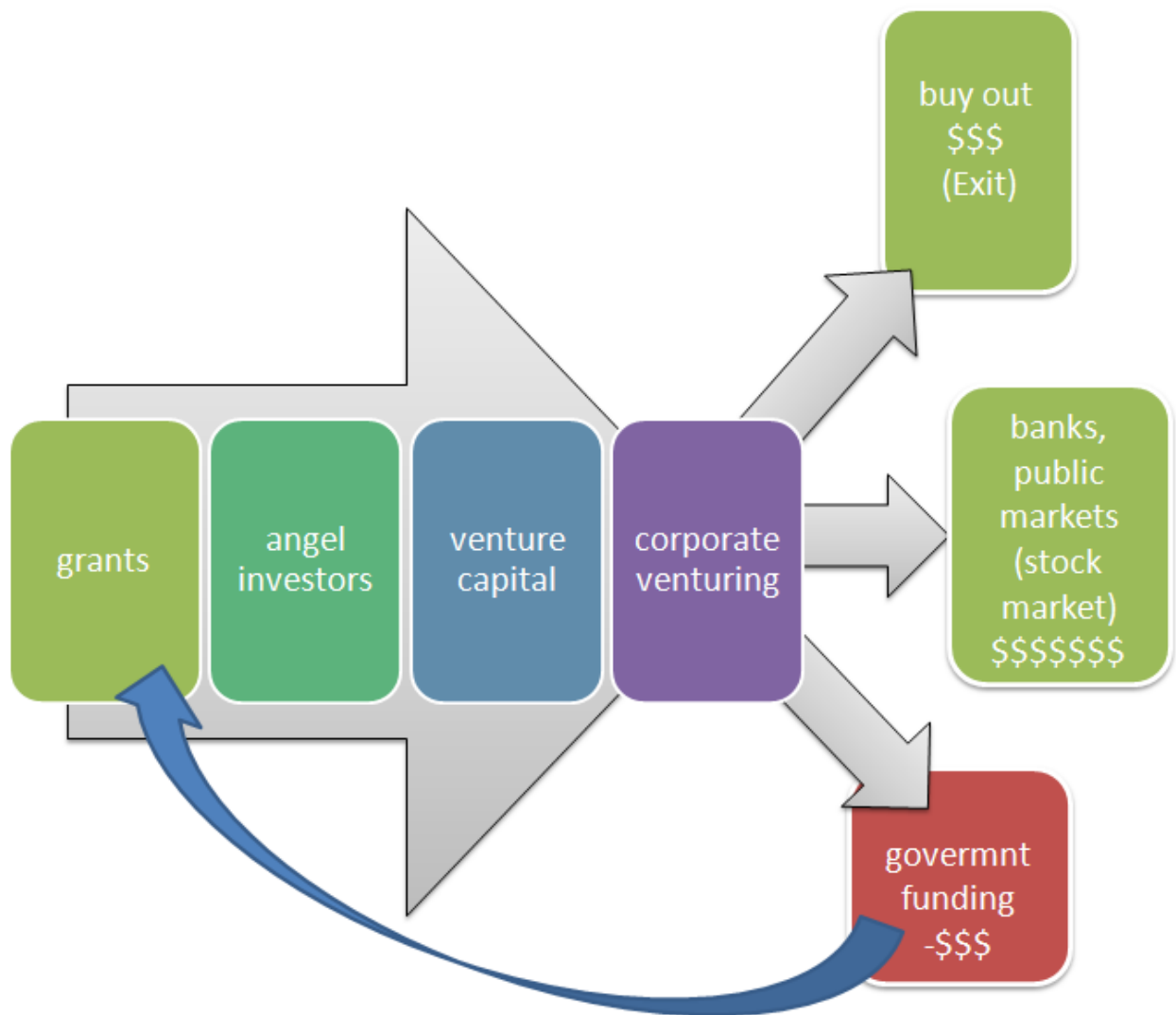


Figure 22 Usual funding stages for high-tech start-ups, with buy out or public floatation as the final stage. The blue arrow is atypical, but observed in some of the proteomic bioinformatics companies in this chapter, such as Oxford Biotherapeutics where government funds are still needed to fund day-to-day operation. Grants at the start are usually at the university stage, but may be in collaboration between a university and company. (Source: author's own summary)

Looking in detail at the business history and case studies, three types of business emerge based on the funding route taken (Table 13), these are:

- Type I: 'ideal' rapid growth, VC-funded, successful exit
- Type II: 'lifestyle' profitable, small, stable, perpetual
- Type III: 'marginal' pre-revenue, need continued support with public grants

Table 13 Funding summary for the proteomic bioinformatics case studies. Oxford Oligosaccharides is taken as an example for comparison with the proteomic bioinformatics companies (the case studies). * - at the grant stage, companies must usually match the funds ‘in kind’ (i.e. by teaching time, resources etc., bank loans or individuals investing).

Type of business	Company	Grants*/ funds from academic projects	Private funds	Angels investors	VC	Corporate venturing	Final or next move
Type I Success – usual funding route	Oxford Oligosaccharides (predecessor of OGS and then OBT)	Yes (Oxford Uni)	Yes (£60m)	No	Yes (Advent Cap. And Euro Ventures)	Yes (Monsanto, Searle)	Floatation (mkt cap £103m)
Type II Stable	Matrix Science	Yes	Yes (shareholders funds)	No	No	No	No change ‘lifestyle’ business
	Proteome Software	No	Yes (director invested initial capital)	No	No	No	Reinvest profits in developing more products
Type III Marginal	Nonlinear Dynamics	Yes (smart awards, northern enterprise, Yes)	Yes (bank loans)	Yes (British Coal)	Yes (NEL)	No	Unknown
	Oxford Bio-therapeutics	Yes	Yes (small amount of shareholders funds)	No	Yes	No	Government funding, reinvest any profits
Unknown, likely marginal Unknown	GeneBio	Yes	?		Yes (index)	?	?
	Sage-N Research	?	?		Yes (Sand Hill)	?	?

Overall, there is no common funding model followed by proteomic bioinformatics ventures. OBT's original company, Oxford Glycosciences was floated on the public markets (market cap: £103m, 1998) then it bought by Celltech for £102m in 2003, thus provides an example of how it can proceed. In proteomic bioinformatics ventures there is no example of type I. The most successful businesses in proteomic bioinformatics (Matrix Science and Proteome Software) are type II: stable, not fast growing, not requiring VC or Angel funding (as described by (Brush *et al.*, 2001)). This is consistent with the market characteristics, being specialised and too small to support the kinds of investments that were seen in the late 1990s and early 2000s: *"Good database search software is not a word processor; it's not a spreadsheet. There aren't millions of customers."* Cottrell, Matrix Science co-founder (Cottrell, 2003). So, the antithesis is that these firms are certainly high-tech, but unlike high-tech seen before they are not suited to VC funding rounds. It appears proteomic bioinformatics companies are a new breed of 'bio-high-tech' firms, unlike IT and unlike typical biotechs.

Type III is the 'marginal' group. They do not demonstrate a clear pattern, but appear to be 'thinking bigger' in terms of their strategy than they should. Nonlinear, for example, had a workforce of a hundred and turnover at £4m, but profits were never higher than Matrix Science, a tiny firm of seven people at its height. The type IIIs show no tangible links with private financial organisations, despite high ambitions for growth – only government backed funds, such as Catapult Growth

private equity fund. They appear to have invested, anticipating growth that the market could not support. The P&Ls quoted for Nonlinear and OBT for 2006-07 are negative, and this was in comparatively good times, when public labs were receiving more funding from the research councils in the UK, before the economic downturn seen in 2008-9. New P&L information is not yet available, but in the short term, the losses may be greater as fewer customers have the ability to buy their products as a result of spending cutbacks.

The case study businesses did not take part in corporate venturing with big pharma or established IT companies, for example (Table 13). This may be because “*an imbalance in the power relationship between a high-tech firm and its network partners makes the high-tech firm vulnerable*” (van der Sijde *et al.*, 2003). There are, however, strategic links and agreements between smaller biotechs and specialist equipment vendors across all the firms examined in the case studies. Indeed, collaboration for small players in this high-tech market is essential, because of the complexity and changing needs of the client: a single entity is unlikely to work flexibly enough to accommodate the changes, but by offering expertise in one specific area, they become an attractive proposition for partnering with other experts to deliver an optimal package. Moreover, other firms may have useful routes to market, which are hard to establish for small start-ups.

Other evidence is provided by the cases (Table 14), for example, the trend for expanding operations into the Far Eastern countries confirms the high-tech nature of proteomic bioinformatics companies. Generally, branding is not an important feature for these companies, because the unique selling points of each software product are sufficient for sales, not usually brand recognition. There are limited options in each niche, because the field is so new, so there is less need to advertise. Matrix Science's Mascot product is perhaps an exception to this, because there are multiple search engines available both public and proprietary, yet Mascot is a trusted 'brand' amongst the proteomics research community.

Table 14 Comparisons between the company case studies

Similarities between case studies	Differences between case studies
Small start-ups	Ambition – Non Linear have highs and lows, Matrix Science is steady low-level growth ‘lifestyle’ business.
Technical individuals as senior management and advisors (scientists, software programmers) not ‘business/commercial’, not high-profile individuals as seen in other industries, such as finance or biotech (such as the biotech ‘champions’ defined in (Markham, 2002))	Matrix Science’s Mascot is the only established ‘brand’ across the board. Most citations in literature.
Major revenue streams are from software licensing, they all have similar product types	Funding routes are diverse (refer to Table 13)
Small niche, high tech firms– cutting edge researchers are their main customer base (industry or academia). Not easy for customers to weigh up the quality of the products before purchase because the field is not well established, so there is less to benchmark and little choice.	Hardware as a bundle in the bioinformatics products – Sage-N product is distinct, offering the Silicon Valley USP
Partnerships, collaborations and agreements with many other companies/institutes. Classic cases of resource sharing seen within collaborative networks – see later subheading regarding the balance between vulnerability and collaboration (van der Sijde <i>et al.</i> , 2003)	
Technology push-based projects – MS technology means there are high-throughput approaches – all the products on offer have emerged to fill the unmet need for complex analysis	
The companies are in the trials, deals and further research stages. Growth ideally needs to be fast enough to be fundable by further VC rounds (Berry, 2002). This is not the case for these firms; they have stagnated.	
Expansion/distribution into Asia-Pacific region, especially Japan and China.	

2.7.2 Specialise for success: the contract research model is recommended

Pharmaceutical companies have removed many core research activities in recent years to try to reduce R&D spending (Prasad, 2004) and CROs have profited from this trend (Sahoo, 2006).

Analysis of the proteomic bioinformatics market suggests that the value chain should be split into individual elements for the range of proteomics research activities for the short-to-mid term (see Figure 23). This process is effectively reorganisation of vertical supply chain relationships (McMillan, 1994).



Figure 23 Suggested concept of fragmentation of the proteomics value chain.

This is a phenomenon where firms shed some the activities they would normally perform in-house, so switch from 'making' to buying. For example, in the 'usual' value chain for an R&D-based pharma organisation, proteomic bioinformatics would have formed part of their own research base, but it is now suggested that small niche firms operate to do this directly for customers, not via an 146 organisation. The net result is that firms reduce headcount and the economy becomes more sophisticated in terms of relationships and specialisms (McMillan, 1994).

The evidence for this recommendation is that proteomic bioinformatics services alone, as provided by Matrix Science for example, appear to be much more commercially successful when they are independent of proteomics data capture facilities. The overheads involved in experimental work are too high; demonstrated by the fruitless land-grab phase and subsequent bust. Too many different areas of expertise had to integrate for success, when each area alone needed to grow and mature independently first.

CROs, such as Quotient BioResearch (a significant collaborator for this EngD) demonstrate the growth in CRO's business. They have rapidly expanded their contract research with pharmaceutical companies, since big pharma now outsources a large proportion of routine experimental work. This trend for outsourcing activities has been seen in many other industries, such as telecoms and

manufacturing, and is already taking place in clinical research for big pharma. The author, with others, such as (Arlington, 2007), forecast outsourcing to increase in scope in pharmaceutical R&D. By 2020, for example it is predicted that 'specialist firms' focussing specifically on discrete areas of drug development, such as testing biological pathways, and proving mechanisms of drug action, will be more prevalent (Arlington, 2007). Proteomics research will most probably be part of this trend.

As the era of the blockbuster drug is over, to replace it there will be a wider array of more targeted medicines. Similarly, big pharma organisations will get smaller and will perform more 'virtual' research through targeted partnerships with smaller expert business and groups in academia, outsourcing large portions of their research operations, such as proteomics research and bioinformatics analysis. Indeed there is already a virtual CEDD (CEEDD) at GSK based on this idea.

R&D efficiency must increase, by these (and/or other) means, because new innovative medicines are desperately needed to cure the chronic illnesses such as diabetes, neurodegeneration, cancer, obesity and heart disease. Costs need to be better managed, by making the right decisions in R&D, and more candidates exhibiting novel mechanisms of action must be found.

These predictions bode well for the aforementioned 'marginal' companies, such as OBT, Nonlinear and Matrix Science. They are suitably placed to soak up the

demand for contract research, as proteomics becomes more developed with SOPs and best practice. A key requirement, also, is the adaptability of these companies: how quickly they can embrace new techniques as the field settles into maturity? Will they be able to get funding to do this?

Moreover, will biomarkers and the opportunity for IP see growth soon? For the moment, it is not likely since the best source of easy-access biomarkers is blood and blood proteins remain problematic, because they are expressed in quantities that vary by ten orders of magnitude; thus, it is difficult to measure interesting, low abundance proteins using current MS-based techniques (Service, 2008b). More focused techniques, such as using N-linked glycopeptides that fish out the interesting proteins, have potential to improve the promise of biomarkers, but overall consensus is that timescales for these developments could be years or even decades.

In summary, the recommendation is to hold off private investment in proteomic bioinformatics companies until the SOPs and platforms become routine. To develop routine technologies, public agencies will provide, and are already providing, the majority of the funding. Big pharma should look further into networking and building relationships with niche CROs, such as the companies discussed in this chapter.

2.8 Conclusion

This chapter has given a detailed account of the development of a new science and the high-tech industry of proteomic bioinformatics.

The business history showed that proteomics was born out of publicly funded research, carried out in the late 1980s, and 1990s. Once high-throughput could be achieved with new MS-related inventions, industry got involved with a view to patenting newly sequenced proteins for developing therapeutics. This was a disaster, as the technology did not generate reproducible results, dogged by technical failings and complexities of the proteome. Divestment ensued in both biotechs and pharma.

The story followed closely the technology hype cycle, where over-enthusiasm for high-throughput proteomics technology was followed by commercial disappointment. The slope of enlightenment and plateau of productivity are still to be reached for high-throughput proteomics, because trust for the technique was lost and the technology has yet to be shown to be reproducible and value-adding.

For now, the market is small and is still immature, supporting only niche companies of a very technical nature. Given the data presented in the cases, these companies

fall into two types: small, profitable and stable companies who have kept headcount to a minimum; and marginal companies, who aimed for higher growth, and now rely on public funds to survive, partly as a result of a previous miscalculation of the size and expected growth of their marketplace.

Analysis of the players in the market suggests that proteomic bioinformatics is indeed a commercially viable activity when executed in a small, specialised company that keeps overheads to a minimum and growth slow and steady. This is in contrast to other equivalent high-tech examples, such as IT, where accelerated growth and several rounds of VC-funding usually ensue.

For corporate pharma, proteomics and associated bioinformatics is unlikely to ever become a large core function, since there will be an increasing trend towards outsourcing technical aspects of R&D to keep costs to a minimum without compromising quality of research. In this future, niche companies will have a refined core competency in the field, so will offer superior quality for less cost than setting up the equivalent research and data analysis infrastructure in-house.

In summary, the author believes that the market for proteomic bioinformatics products and services will not grow significantly until standards are developed, and the value added by the technologies can be clearly demonstrated to new investors.

As a result of the problems in the past, investment in a high-throughput approach

may be harder to obtain, but the fact remains that proteins are the molecular machines of the cell. Only by understanding these can a radical breakthrough be made to improve drug design processes and deliver understanding and targeting of the most elusive diseases.

Review of public repositories for proteomics

"To hoard is human; to share, divine."

From (Wells *et al.*, 2008)

3.1 Summary

Since the withdrawal of high throughput proteomics at GSK (2005-6), and given that there are no in-house labs producing proteomic MS data at Cranfield University, a major source of data for the research in this EngD project was public data. This is data deposited into public proteomic repositories on the internet, usually the product of publicly-funded research projects and may be downloaded by anyone for free.

The problem with using public data, however, is that new data appears frequently but is not usually announced through traditional routes, such as research papers, and developments in the repositories themselves are frequent. For this reason, there is an urgent need for a single document detailing all available resources, so researchers can leverage the most value from the systems and data available to them. Thus, for both practical purposes and to add to state of the art, this chapter presents a review of the major public data repositories, the data they contain and the pipelines that populate them.

3.2 Introduction

3.2.1 Definition of a public proteomics repository

As explained in the introduction public proteomics repositories can store and disseminate proteomic MS datasets. In this way they are distinct from the freely available laboratory information management systems (LIMS) such as PROTEIOS (Garden *et al.*, 2005), PRIME⁶⁹, YASSDB (Thomsen *et al.*, 2007), Labkey.org's CPAS (Rauch *et al.*, 2006) and the Proteomics Experiment Data Repository (Taylor *et al.*, 2003). This is because LIMS systems typically store much more diverse data such as gel images, plate barcodes and protocol information, and are primarily intended for local data analysis and archiving, rather than for public data sharing over the internet; however, some do also facilitate secure data sharing across geographically distant collaborating groups. As mentioned in the Introduction, proteomics repositories are also distinct from protein databases, because they store data pertaining to MS/MS experiments, not data pertaining to proteins *per se*. Commercial products for proteomics data capture and storage are not described here, because they are not relevant for this EngD project.

3.2.2 Benefits of public repositories

As proteomic MS has increased in throughput, so has the demand to catalogue the increasing number of peptides and proteins observed by this technique. As in other 'omics' fields, this brings obvious scientific benefits such as sharing of results and

⁶⁹ <http://prime.proteome.med.umich.edu>

prevention of unnecessary repetition, but also provides technical insights, such as the ability to compare proteome coverage between different laboratories, or between different proteomic platforms (Figure 24).

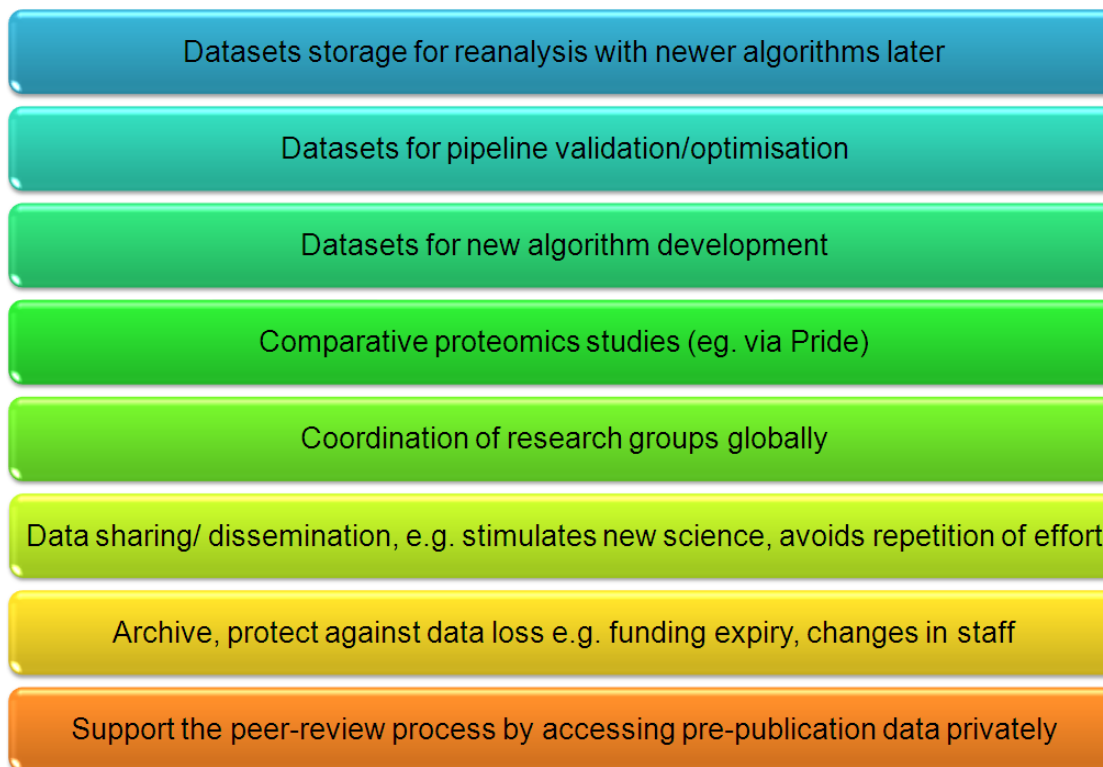


Figure 24 A summary of the benefits and potential uses of public proteomic MS repositories

As well as offering direct benefits, proteomic data repositories have also catalysed developments in other areas of proteomics research (Figure 25); for instance, the availability of large volumes of data - only possible by combining efforts from many labs - means research can now be performed and biological conclusions drawn which otherwise would have been impossible. A specific example is PTPs: only when redundant data is available can PTPs be identified, and the benefits of

knowledge of PTPs be gleaned, such as design of PTP-based search engines optimised for speed, like X!P³ (Craig *et al.*, 2005), or software suites for MRM transition design, like in TIQAM (Lange *et al.*, 2008). Furthermore, with access to large amounts of data the limitations of the MS method *per se* are brought to light, for example redundancy in the peptide identification data can effectively demonstrate the limited range of visibility of peptides using current techniques (Nesvizhskii *et al.*, 2007) - stimulating the community to look for improved alternatives for detecting the elusive peptides.

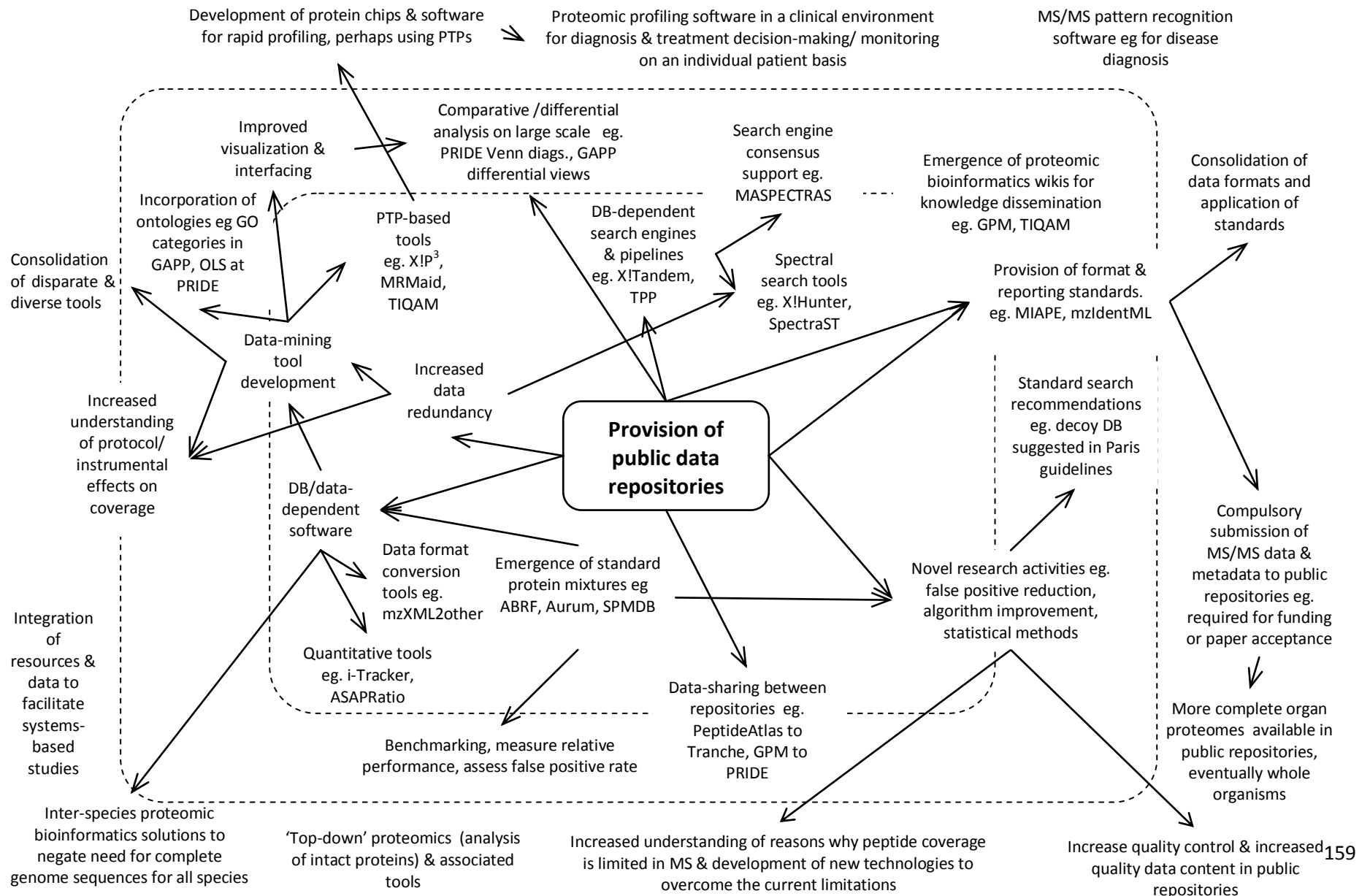


Figure 25 The emergence of public proteomic data repositories has stimulated the development of a huge array of public bioinformatics resources, including pipelines and diverse data analysis tools. The concentric dashed ellipses show the progress of the field, with the central ellipse showing the completed areas of work, the next showing areas of work in progress, to the outermost area, which describes anticipated future developments in the field. Text which overlaps a boundary means that the work transcends both areas. (Source: author's own summary)

Another advantage of availability of public datasets is that advanced data-mining and visualisation tools can be developed. These programs, which sit at the 'front end' of some public data repositories, can highlight important trends in data that would otherwise remain hidden; for example, tools to collate and display differentially expressed proteins, peptides and PTMs, or data-mining programs that use Gene Ontology (Ashburner *et al.*, 2000) categories to compare shifts in cell or molecular functionality across datasets. Such tools can be used by researchers to analyse their own data, possibly in the context of data from other groups, and as the amount of data in the repositories increases, so does the possibility of new discoveries being made purely using existing public data.

3.3 An overview of the public repositories

The main public repositories are the Proteomics IDentifications database (PRIDE) (Jones *et al.*, 2006), the Global Proteome Machine database (GPMDB) (Craig *et al.*, 2004), PeptideAtlas (Desiere *et al.*, 2006), Tranche at ProteomeCommons⁷⁰, the Genome Annotating Proteomic Pipeline (GAPP) (Shadforth *et al.*, 2006), and Human Proteinpedia (HPP) (Mathivanan *et al.*, 2008).

⁷⁰ Falkner, J. A., Andrews, P. C., HUPPO Conference 2006, Long Beach USA, poster presentation

Smaller scale repositories include the Max-Planck Unified Proteome Database (MAPU) (Zhang *et al.*, 2007), PepSeeker (McLaughlin *et al.*, 2006), and SwedCAD (Falth *et al.*, 2007). And even more specialist offerings include the Yeast Resource Center Public Data Repository, the BiblioSpec Library (Frewen and MacCoss, 2007), the Open Proteomics Database, the Proteomics Data Center at the Resource Center for Biodefense Proteome Research (Zhang *et al.*, 2008), SWISS-2DPAGE (Hoogland *et al.*, 2004), and Biodemo; these are not covered in this chapter.

The main repositories are now described briefly, followed by more detailed information of the features for the six databases deemed relevant for this EngD, namely: PRIDE, GPMDB, PeptideAtlas, Tranche, GAPP and Human Proteinpedia.

3.3.1 PRIDE



PRIDE⁷¹ is not limited to identifications; it also includes peak list data for download, journal article links and associated tools, including data-mining, visualisation and ontology-assisted data conversion tools (Jones *et al.*, 2008b). With close links to HUPO PSI, PRIDE aims to be compliant with agreed community standards, in terms of reporting (MIAPE) and standard data formats (mzML and mzIdentML), as soon as they become available. PRIDE does not include an analysis pipeline and as such stores data from any appropriate MS/MS analysis workflow. This repository has a facility for pre-publication data storage to assist the peer review process. Notable

⁷¹ <http://www.ebi.ac.uk/pride/>

datasets in PRIDE include the acid mine drainage extract, HUPO liver (HLPP), HUPO plasma proteome (HPPP), HUPO brain and human CSF, and the Cellzome dataset.

3.3.2 GPMDB



GPMDB⁷² celebrated its 50,000,000th peptide identification on 16th May, 2008. It was created by Ron Beavis and co-workers, and was originally designed as a web-interface for the X!Tandem Spectrum Modeler search engine (Craig *et al.*, 2004), but it has been developed significantly, now allowing comparison between experimental results and the best results that have been previously observed by other scientists. It is the first repository to apply analytics tools to map the number of visits, and has recently developed new features: for example, it has expanded the number of eukaryotic species supported (as NCBI builds) and shifted to collecting annotated spectral files for improved searches, as well as a new compressed data format (called Common, .cmn). A peer-to-peer grid computing system, called Tornado, has also been released which speeds up searches tenfold (for human, mouse and rat) by determining which server on the grid is least busy and sending the search to X!Tandem there. In addition, there are new ways to access and view the data, such as a new view that allows protein lists to be explored by chromosome number⁷³ or those derived from mitochondria or transposons, also there is an MRM

⁷² <http://gpmdb.rockefeller.edu/>

⁷³ http://gpmdb.thegpm.org/go/index_chr.html

worksheet that uses consensus spectra to select transitions for MRM (Walsh *et al.*, 2009).

3.3.3 PeptideAtlas



PeptideAtlas⁷⁴ is a project of ISB, Seattle, and as such its creators place an emphasis on its application to systems biology research. It is described as a “*platform to select and validate MS targets*” (Deutsch *et al.*, 2008) and is separated into ‘builds’, which represent all peptides mapped to a single reference Ensembl genome. This allows protein identifications to be viewed from within the Ensembl browser as a DAS track (Dowell *et al.*, 2001, Shadforth and Bessant, 2006). Current species builds include human, human plasma, *Drosophila*, *Drosophila* Phosphopeptide, yeast, mouse, halobacterium and *Streptococcus pyogenes*. The Ensembl and IPI accession numbers are supported and PeptideAtlas is also home to the Human Plasma Proteome Project Data Central (Omenn *et al.*, 2005) with various links to project-specific articles and identification data.

MS data is made available as peak lists by submitting laboratories, from which TPP (Trans Proteomic Pipeline) (Keller *et al.*, 2005) extracts peptide IDs to populate the SBEAMS (Systems Biology Experiment Analysis Management System) proteomic DB module. Furthermore, PeptideAtlas has released a raw data repository for MS/MS dataset posting, acting as a data provider for others, including the spectrum library at NIST and the PepSeeker database (McLaughlin *et al.*, 2006).

⁷⁴ <http://www.peptideatlas.org/>

To derive a quantitative perspective on protein expression, for generation of system models and simulations, PeptideAtlas has developed an MRM transition prediction tool, TIQAM (Lange *et al.*, 2008), and a database of yeast MRM transitions, MRMAAtlas (Picotti *et al.*, 2008).

3.3.4 Tranche at ProteomeCommons



Tranche

Tranche⁷⁵ aims to solve the problem of data sharing in proteomics by supporting transfer and dissemination of very large datasets in a secure fashion across the internet. It has huge volumes of distributed disk space for backing-up of proteomics datasets, facilitating long term data storage to ensure data is not lost through changes in staff or funding. It also, like PRIDE, has a facility for pre-publication (private) data storage. It is a storage platform not a relational database, so does not provide powerful querying functionality. The system had 5,502 projects and 11.1 million individual files corresponding to 3.1 terabytes on June 26th, 2008. Notable datasets include The National Cancer Institute (NCI) Mouse Proteomics Technologies Initiative (MPTI) project and Kislinger and co-workers' ascites study into ovarian carcinoma biomarkers (Gortzak-Uzan *et al.*, 2008). PeptideAtlas data repository is also mirrored here. Tranche has an average of 70 website visits per day (reported on April 22nd, 2008).

⁷⁵ <http://tranche.proteomecommons.org/>

3.3.5 GAPP



GAPP⁷⁶ is a data analysis pipeline, the results of which are stored in GAPP DB. Like PeptideAtlas identifications can be viewed as an Ensembl DAS track (Shadforth and Bessant, 2006). Much of the data in GAPP is taken from other repositories, and has been reanalysed to allow direct comparison between different datasets.

GAPP has undergone significant developments since creation on a previous EngD project, and in its current state it is a critical element required for the novel research work presented in this thesis. New graphical visualisations and data-mining functionality have been developed for GAPP by others in the Bioinformatics Group, as well as simplification of data submission using data-entry forms.

3.3.6 Human Proteinpedia

HUMAN PROTEINPEDIA HPP⁷⁷ is a repository of diverse proteomic datasets including data from MS/MS, co-immunoprecipitation MS/MS, immunohistochemistry, yeast two-hybrid and other platforms. It is the best repository in terms of consistent reporting of all necessary metadata to reprocess the raw datasets accurately; routinely reporting mass and fragment tolerances, for example. Its data content is steadily increasing, with 71 labs regularly submitting.

⁷⁶ <http://www.gapp.info>

⁷⁷ <http://www.humanproteinpedia.org/>

The HPP repository is complemented by The Human Protein Reference database (HPRD) (Mishra *et al.*, 2006): a database of literature-derived information, compiled at the level of the individual protein. HPRD shows, for example, characterised protein domains, PTMs and interactions, and allows users to explore the meaning of their data, submitted via HPP. Each protein in HPRD is annotated with peptide information and the sample of origin (and where relevant as an HPP 'HuPA' identifier).

3.3.7 MAPU



MAPU 2.0 ⁷⁸ a database of organellar, cellular, tissue

and body fluid proteomes was released in 2006 by the Max-Planck-Institute of Biochemistry in Martinsried, Germany (Zhang *et al.*, 2007). The system is a family of discrete proteome databases, which its creators believe will eventually provide reference proteomes for biomarker discovery studies. It has a notebook-style design with tabs to navigate around the site, and like PeptideAtlas and GAPP, it has genome annotating functionality viewed as a DAS source. Furthermore, transcript annotations are also accessible via a graphical chromosome view, similar to GPMDB. The proteomes available at MAPU 2.0 include mouse adipocyte and liver, and for human, there are body fluid proteomes: urine, tears and seminal fluid. MAPU 1.0 also includes the human plasma and cerebrospinal fluid proteomes.

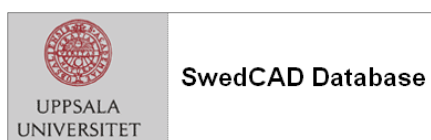
⁷⁸ <http://www.mapuproteome.com>

3.3.8 PepSeeker

The logo for PepSeeker, featuring the word "PEPSEEKER" in white, bold, uppercase letters on a dark blue rectangular background.

PepSeeker⁷⁹, unlike other repositories, aims to increase the understanding of peptide fragmentation chemistries and the effect of peptide sequence on visibility in MS (McLaughlin *et al.*, 2006). It is led by Simon Hubbard at the University of Manchester, UK and it stores peptide identifications and corresponding fragment ion details used to identify that amino acid sequence. The argument for such a system is that the peptide sequence composition determines the presence or absence of certain ions in the resulting spectrum, so by harnessing this information, development of more sophisticated peptide identification algorithms should be possible in the future. Improvements to query functionality have been made (now via Biomart), and it has contains a 'gold' database instance comprising only the highest scoring peptides.

3.3.9 SwedCAD and SwedECD



SwedCAD⁸⁰ and SwedECD⁸¹ are databases of high resolution, high mass accuracy MS/MS spectra derived from collision associated dissociation (CAD) and electron capture dissociation (ECD) MS, respectively. One aim of the repositories is to provide reference spectra for use as search databases for peptide identification, since all

⁷⁹ <http://www.nwsr.manchester.ac.uk/cgi-bin/pepseeker/pepseek.pl?Peptide=1>

⁸⁰ <http://www.bmms.uu.se/CAD/>

⁸¹ <http://www.bmms.uu.se/CAD/indexECD.html>

datasets may be downloaded for local use. Arguably, the main aim, however, is to provide a repository that can offer unique insight into fragmentation pattern aetiology, such as exploring the phenomenon of neutral loss, looking at the frequency and nature of missed cleavages, and the effect of certain motifs, such as terminal 'RR' (Falth *et al.*, 2007). The systems are able to do this because they provide both CAD and ECD spectra for the same samples, namely human milk, lysates of human cell lines (K562 and A-431) and *E.coli* proteins.

3.4 Data upload, download and format support

New standard data formats for MS are available, as outlined in Chapter 1, but they must become “the norm” in order to be useful to the community. To achieve this, the analysis tools and databases must support them, thus encouraging their use. In the short term, however, diverse data formats remain in existence so knowing which system supports which formats is necessary – thus is described here. Long-term, it is envisaged that the standards will prevail across all public MS resources.

3.4.1 PRIDE

PRIDE supports private data submission, generating anonymous login details to grant access to the uploaded dataset. This anonymous login can be sent to reviewers, providing confidential access to the details of the proteomics experiment supporting the manuscript before publication. The Proteome Harvest PRIDE Submission Spreadsheet is available for small-scale submissions of data to PRIDE (Jones *et al.*, 2008b) supporting conversion to PRIDE 2.1 XML and applies an ontology look-up

service (OLS) (Côté *et al.*, 2006) to apply controlled vocabularies to data annotations. Data may also be submitted via the MASPECTRAS pipeline (Hartler *et al.*, 2007), or via PrideWizard (Siepen *et al.*, 2007) (for Mascot files), which both create PRIDE XML output. The method of choice, however, is the new PRIDE Converter (Barsnes *et al.*, 2009), which can be used for large or small datasets, accepts various input formats, and has a 'point-and-click' graphical interface.

Once registered, users may submit data to PRIDE as individual XML files (PRIDE XML or mzData/mzML), or as a zip archive of multiple files. Submitting data in this way can be impersonal; however there is access to a curator at PRIDE if submission is difficult. For bulk data submissions, PRIDE also supports secure FTP upload of datasets.

In PRIDE's case, public mzData/mzML and PRIDE XML datasets may be downloaded. The files are downloaded as zip files directly from the website or alternatively the FTP server⁸². The only exception to this is where experiments have been submitted without MS/MS spectra, in which case mzData/mzML is not available.

⁸² <ftp://ftp.ebi.ac.uk/pub/databases/pride/>

3.4.2 GPMDB

Adding data to GPMDB is possible via the search pages in public or restricted mode. To submit data for analysis and storage a species (referred to as a 'boutique') is selected and the details for the search are entered on the Tornado search form. Only the top 50 most intense ions are applied from the peak lists (Walsh *et al.*, 2009).

GPMDB's new compressed format is Common (.cmn), and there is a tool provided to compress/decompress MS/MS data files (such as .mgf, mzXML, mzData, .dta and .pkl) to/from .cmn. MS/MS files may still be submitted in more familiar formats, however .cmn files have the scope to support analysis of files that would previously have been too large to submit. Furthermore, .cmn files can be made available to the data archive⁸³. GPM also provides protein lists derived from specific tissues, as part of the Normal Clinical Tissue Alliance⁸⁴. Both processed data and the raw MS/MS files are available, the former via GPMDB interfaces in the usual way, and the latter via an FTP site⁸⁵.

3.4.3 PeptideAtlas

MS/MS spectra are accepted in either native (.raw) format or in mzXML, with the latter being preferred. There is a web application for data submission, and to access it the administrators must be contacted. As part of PeptideAtlas' pipeline, the Trans-Proteomic Pipeline (TPP), there are several free tools available. For data submission,

⁸³ <ftp://ftp.thegpm.org/data/msms>

⁸⁴ http://wiki.thegpm.org/wiki/Normal_Clinical_Tissue_Alliance

⁸⁵ ftp://ftp.thegpm.org/projects/ncta/release_1/data

for example, there is a program⁸⁶ to convert .wiff⁸⁷ format files to mzXML. If desired, a date can be entered for when the raw data becomes available to the public. 'Minimum public access' submissions still appear on the site with the institution name, the contact, date, organism and cell type, but data is only available when it becomes public. To download unprocessed datasets and lists of identifications, there is a Data Repository⁸⁸ - by clicking on the zip files they may be saved locally. Accompanying journal paper links are also provided.

3.4.4 Tranche

Tranche is based on peer-to-peer data sharing. To upload data, a Tranche account must be requested and any data format is accepted. The administrators also offer a data submission service with data on USB hard drives. An easy way to download data is via the website⁸⁹, where there is a long list of the available datasets. The free Java downloader package manages the download, and also executes the security check for private datasets. Tranche applies industry standard encryption protocols for security. In the current system, data annotations can be made by anyone uploading data - a separate username and password is not required.

3.4.5 GAPP

As GAPP is primarily an analysis pipeline rather than a repository, it accepts mass spectra not identifications, and does not provide data for download. To submit data,

⁸⁶ <http://tools.proteomecenter.org/wiki/index.php?title=Software:mzWiff>

⁸⁷ Proprietary raw data format used by Applied Biosystems

⁸⁸ <http://www.peptideatlas.org/repository/>

⁸⁹ <http://www.proteomecommons.org/data.jsp>

users must register and login, then create a dataset-specific 'data profile', which is an electronic form for metadata. Data may be submitted in .mgf, .pkl or mzXML format, and can be submitted privately. Privately submitted data is not visible to any other users at any time, including via the data-mining tool, MR Maid, described later in this thesis. GAPP DB (the current version) captures and analyses only the 100 most intense ions.

3.4.6 HPP

A comprehensive list of meta-annotations is required to submit data to HPP. The submission process exploits ontologies where necessary and guides the submitter by providing drop down menus. Format is not specified, with currently available datasets including .pkl, .mgf, mzXML and .raw. To submit, users must first register and login. All data is visible to the public and is stored in triplicate across 16 servers. HPP employs the Tranche Java application for download of raw datasets and identifications.

3.5 *Data-mining and visualisation*

3.5.1 PRIDE

There are two ways to mine data in PRIDE: by 'browsing experiments' (Figure 26(a)), or by querying with Biomart. The former is a 'flat file' table view with hyperlinks, whereas the latter is based on the programmatic web service framework (Jones *et al.*, 2008b) offering flexible searching and quick retrieval of user-specified relevant data. Half of all downloads from PRIDE are via Biomart.

Since PRIDE accepts identifications that are derived from searching different protein databases, querying across all datasets exploits the Protein Identifier Cross-Reference Service (PICR) (Côté *et al.*, 2007), which is cross-referencing tool to map identifiers across 60 different databases. Thus, PRIDE may be queried by the users' accession numbering system of choice. Furthermore, PRIDE has been incorporated into EBI's new metasearch, 'EB-Eye' - a search tool that collates data from multiple EBI resources.

PRIDE also has graphical *de novo* sequencing spectrum views (Figure 26(b)) and a tool to dynamically generate Venn diagrams to compare identifications from up to three datasets (Figure 26(c)).

(a)

Search Summary View

Select an Experiment (The results will remain restricted according to your original search)
Search filtered on: Experiments Browsed By CV Term , with parameters Bartschhoff:2007-1

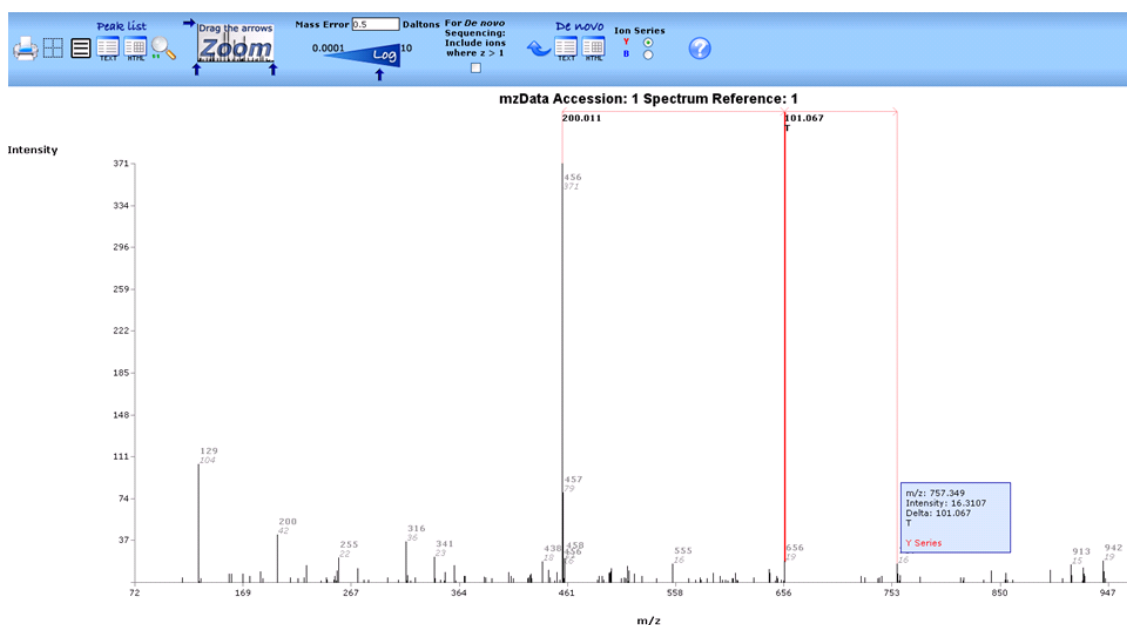
[View Instructions](#)

This Table Describes 734 Experiments.
Sorting has been disabled because the table exceeds 400 rows.

[Compare Experiments](#)

Accession	Title	Species	Tissue	Cell Type	GO Term	Disease	Protein Count	Peptide Count	Spectra Count	Retrieve Details (View in web browser or download as XML file)	Compare Protein Identification Sets	Select Reference Experiment
2445	Cellzome_Abi_inhibitors_NatureBiotechnology_exp1	Homo sapiens (Human)	HeLa cell				6	228	228	View Download	<input type="checkbox"/>	
2446	Cellzome_Abi_inhibitors_NatureBiotechnology_exp2	Homo sapiens (Human)	HeLa cell				3	57	57	View Download	<input type="checkbox"/>	
2447	Cellzome_Abi_inhibitors_NatureBiotechnology_exp3	Homo sapiens (Human)	HeLa cell				2	161	161	View Download	<input type="checkbox"/>	
2448	Cellzome_Abi_inhibitors_NatureBiotechnology_exp4	Homo sapiens (Human)	HeLa cell				3	174	174	View Download	<input type="checkbox"/>	
2449	Cellzome_Abi_inhibitors_NatureBiotechnology_exp5	Homo sapiens (Human)	HeLa cell				5	126	126	View Download	<input type="checkbox"/>	
2450	Cellzome_Abi_inhibitors_NatureBiotechnology_exp6	Homo sapiens (Human)	HeLa cell				7	337	337	View Download	<input type="checkbox"/>	
2451	Cellzome_Abi_inhibitors_NatureBiotechnology_exp7	Homo sapiens (Human)	HeLa cell				5	231	231	View Download	<input type="checkbox"/>	
2452	Cellzome_Abi_inhibitors_NatureBiotechnology_exp8	Homo sapiens (Human)	HeLa cell				6	260	260	View Download	<input type="checkbox"/>	
2453	Cellzome_Abi_inhibitors_NatureBiotechnology_exp9	Homo sapiens (Human)	HeLa cell				5	147	147	View Download	<input type="checkbox"/>	

(b)



(c)

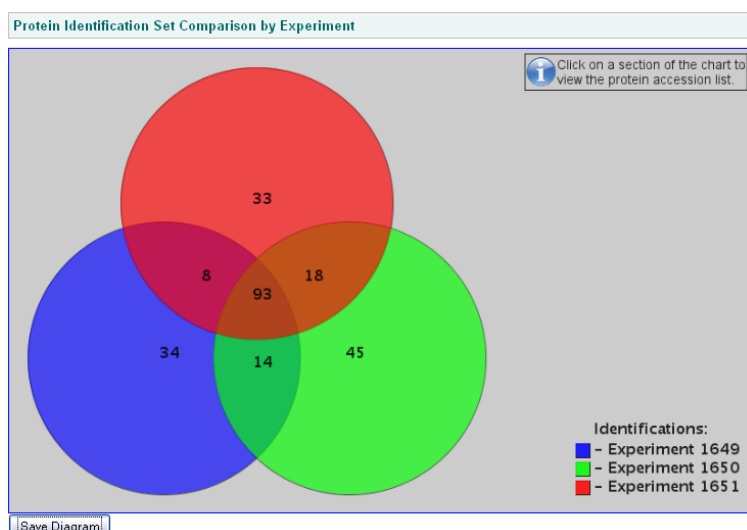


Figure 26 PRIDE data visualisation options (a) dataset download page; (b) manual *de novo* sequencing tool; (c) Venn diagram for comparing identifications across experiments

3.5.2 GPMDB

For each protein entry in GPMDB, all the experiments in which the protein was identified are listed with a schematic representation of the protein sequence with the individual red 'peptide blocks' for observed peptides, and green blocks for regions that are predicted to be difficult to observe in MS/MS (Figure 27).

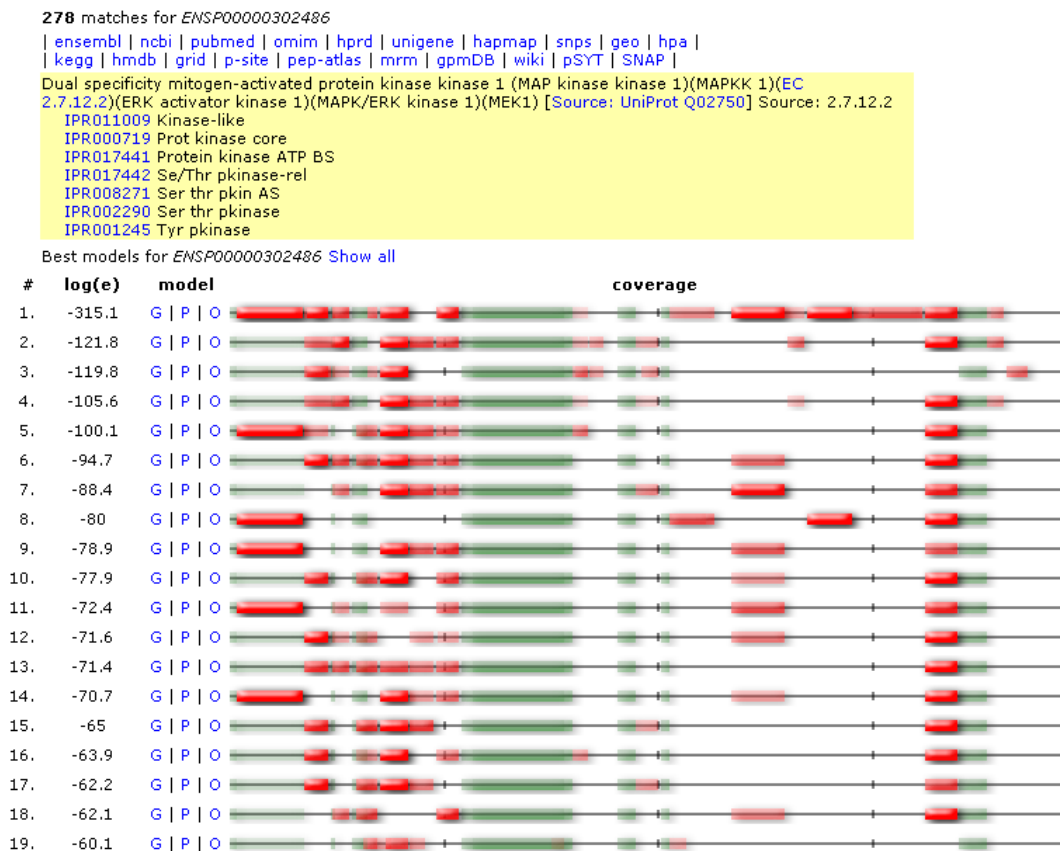


Figure 27 GPMDB's experiment view with assigned peptides shown as blocks along the protein sequence. Each row represents a single experiment. Red blocks are PTPs, green are predicted to be not visible in MS

Statistical significance of identifications is also illustrated graphically, by the shade of red, and there are three possible ways to view identified proteins: gene view (G), protein and observed peptide sequence view (P) and the X!Tandem view (X). X is a

collapsible view of the algorithm's output, which breaks down to the details, including the x,y-coordinates of the original peak list. Graphics are implemented in SVG.

As in PRIDE, different protein database accession systems can be searched, however GPMDB has implemented automatic conversion of identifiers (for Ensembl, IPI, HGNC or MGI gene symbols, NCBI genes and Swiss-Prot/Uniprot) without requiring repeated searches. The 'note keywords' search allows querying of submitter's notes (that accompanied the uploaded datasets), and there is batch querying where a list of peptide sequences is queried. In addition, when searches have been performed against the Ensembl database (for human, mouse, rat or yeast) the results may be viewed as a KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway by sorting proteins into metabolic pathway categories using ontologies.

3.5.3 PeptideAtlas

Peptide and protein information may be browsed, where each peptide entry has a list of properties (like pI, sequence, accession number and SSRCalc (Krokhin *et al.*, 2004)-derived hydrophobicity) (Figure 28). Each protein entry has a genome view, which shows observed peptides and predicted domains, such as membrane spanning regions determined by TMHMM (Krogh *et al.*, 2001). As in GPMDB's view, peptide regions that are predicted to be unlikely to be visible in MS/MS are highlighted (Figure 28).

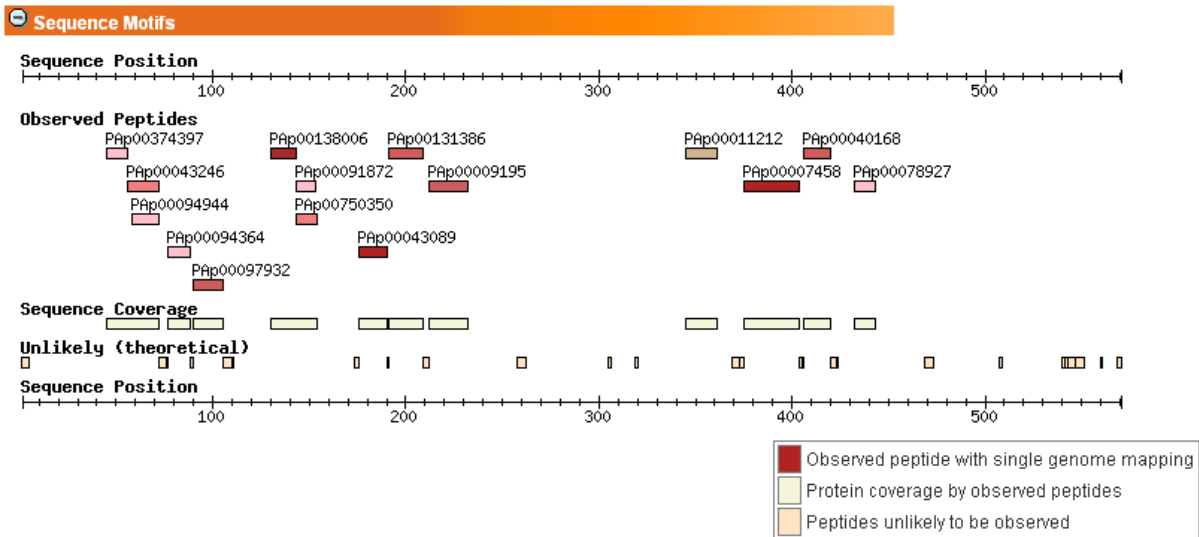


Figure 28 PeptideAtlas maps peptides on to the genome using a DAS track.

PeptideAtlas has the Cytoscape (Shannon *et al.*, 2003) plug-in linked to each protein entry, which allows peptides to be viewed as a network with associated proteins. The proteotypic score reflects the likelihood that a peptide is proteotypic, this is also shown in the protein view. Hyperlinks to external resources, like Ensembl or IPI, are also provided.

3.5.4 Tranche

The data available via Tranche is listed on a single page (linked via the 'data panel' shown in Figure 29). A simple search form is available for specifying the search criteria, such as, project status, journal, and researcher name.

Our Tranche Network

- About
- User Guide
- Admin Guide
- FAQ
- Collaborations
- Downloads
- Servers
- More Info @ TrancheProject.org

Launch Tranche!

Trouble launching Tranche?

Data (8,236)

- + PeptideAtlas repository PAe000030_prot_4130_200909281049.tar.gz
- + PeptideAtlas repository PAe000030_Search_Results_4130_200909281030.tar.gz
- + PeptideAtlas repository PAe000030_mzXML_200909281026.tar.gz
- + PeptideAtlas repository PAe000277_README
- + PeptideAtlas repository PAe000277_prot_4137_200909270339.tar.gz
- + PeptideAtlas repository PAe000277_Search_Results_4137_200909270336.tar.gz
- + PeptideAtlas repository PAe000277_mzXML_200909270333.tar.gz
- + PeptideAtlas repository PAe000073_prot_4132_200909270331.tar.gz
- + PeptideAtlas repository PAe000172_README
- + PeptideAtlas repository PAe000172_prot_284_200909270311.tar.gz

What is Tranche?

The Tranche Project is a free and open source file sharing tool that enables collections of computers to easily share and cite scientific data sets. [More...](#)

Statistics

Server Count	17
Used Disk Space	20.7 TB
Total Disk Space	60.1 TB
Data (uncompressed)	10.8 TB
Data (compressed)	6.9 TB
Data Files	12,641,702
Avg Download Speed	1.3 MB/s
Avg Upload Speed	712.9 KB/s

Figure 29 Homepage for Tranche at ProteomeCommons. Latest datasets are shown in the centre.

3.5.5 GAPP

GAPP now provides three options for querying: by experiment, by protein, or by 'differential view'. The experiment view lists all experiments processed and stored in GAPP by displaying headline information about each. By selecting a single experiment, metadata for the experiment and lists of identified peptides and proteins are displayed in a collapsible view. The protein view allows users to search for a protein of interest by Ensembl accession number and displays a breakdown of information for this protein, including peptide coverage, GO categories, number of experiments in which it was seen and PTMs, where found.

The differential view allows protein expression to be compared between experiments in a table. These experiments may be selected manually, or according to metadata (e.g. tissue type, disease state or instrument type). As the list of proteins in

such a comparison can be large, proteins can be filtered according to GO category (Figure 30(a)), and a pie chart (Figure 30(b)) illustrating the breakdown of protein identifications is generated dynamically for the selected experiments.



Proteins/Experiments	5	6	7
ENSG00000012223			
ENSG00000051415			
ENSG00000090382			
ENSG00000090920			
ENSG00000091513			
ENSG00000092820			
ENSG00000101441			
ENSG00000101443			
ENSG00000108518			
ENSG00000111215			
ENSG00000117983			
ENSG00000120885			
ENSG00000125999			
ENSG00000126550			
ENSG00000131686			
ENSG00000132465			
ENSG00000135046			
ENSG00000148180			
ENSG00000148346			
ENSG00000149021			
ENSG00000159763			
ENSG00000160180			

(a)



- (28%) cellular process
- (9%) developmental process
- (11%) biological regulation
- (15%) response to stimulus
- (10%) metabolic process
- (5%) localization
- (7%) immune system process
- (2%) multi-organism process
- (8%) multicellular organismal process
- (2%) biological adhesion
- (1%) reproduction
- (0%) cell killing
- (0%) growth

(b)

Figure 30 GAPP's differential view shows (a) a comparison of the protein content between experiments 5, 6 and 7, and (b) the proteins found in the experiments broken down into GO 'biological process' categories.

3.5.6 HPP

There are three search avenues: genes/proteins, annotations and MS platform. Ontologies are implemented and the user may select from lists of agreed terms using the query form. For gene or protein searches, entries that identified the query gene or protein are listed, and under each contributor's and experiment information there are the peptides that contributed to the identification with associated modifications, charge state, precursor mass and HuPA identifier. Complete datasets may be downloaded with corresponding metadata (Figure 31).







	Human Proteinpedia Accession Number	Description	Journal Name	Annotation Category	Experimental Platform	Download Data
1.	HuPA_00001	Brain proteome analysis		Tissue expression: Brain	Mass spectrometry	Download
2.	HuPA_00002	Plasma proteome analysis	Unpublished	Tissue expression: Blood plasma	Mass spectrometry	Download
3.	HuPA_00003	B Cell proteome analysis		Tissue expression: B Cell	Mass spectrometry	Download
4.	HuPA_00004	Saliva proteome analysis	Unpublished	Tissue expression: Saliva	Mass spectrometry	Download
5.	HuPA_00005	Plasma proteome analysis	Unpublished	Tissue expression: Blood plasma	Mass spectrometry	Download
6.	HuPA_00006	Co-immunoprecipitation based protein-protein interaction analysis		Protein-protein interaction	Co-immunoprecipitation based protein-protein interaction	Download
7.	HuPA_00007	Platelet subproteome analysis	Unpublished	Tissue expression: Platelet Subcellular localization	Mass spectrometry	Download
8.	HuPA_00008	Serum proteome analysis		Tissue expression: Serum	Mass spectrometry	Download
9.	HuPA_00009	Serum proteome analysis		Tissue expression: Serum	Mass spectrometry	Download
10.	HuPA_00010	PARK7 brain expression analysis	Unpublished	Tissue expression: Brain	Western blotting	Download
11.	HuPA_00012	Her2/neu tyrosine phosphoproteome analysis		Cell line expression: NIH3T3 Post-translational modifications	Mass spectrometry	Download

Figure 31 Human Proteinpedia's download page. Raw datasets, protein identifications and meta-annotations may be downloaded for some experiments, for other just identifications and metadata are available.

Although there is comprehensive search functionality at HPP, arguably more powerful biological querying may be performed using the reference database, HPRD (Mishra *et al.*, 2006), which is directly linked to HPP and also contains peptide

sequences derived from the PeptideAtlas and PRIDE repositories. HPRD can be queried using gene symbols or various accession numbers including RefSeq, OMIM, Swiss-Prot, HPRD and Entrez Genes. Several different parameters may be queried simultaneously as well as via a BLAST search tool. There are also links to curated pathway information and visualisations.

3.6 Data content of repositories is varied

Since each repository fits a particular niche and the developers have different collaborators, the data present in each system can vary. This section highlights the differences in data volume (Table 15).

Table 15 A summary of data content of the major public repositories. Values were reported on 26th June 2008, except PeptideAtlas where values were taken from (Deutsch *et al.*, 2008), and PRIDE where values were taken from a lecture (Phil Jones, 4th December 2008, Cranfield University). a) denotes that the value is not readily obtainable on the website or recent paper; b) referred to as 'projects'; c) these values are based on the gene level, with values indicating distinct identifications at the gene level ; d) high confidence proteins; e) two *Salmonella typhimurium* datasets are counted so identifications may overlap; f) includes post-translationally modified and different splice variant peptides, for the other values in this column it is assumed that they are counted but not stated explicitly in the data sources.

Repository	Species	No. spectra	No. experiments	Protein identifications	All peptides	Distinct peptides
PRIDE	all	16,564,434	7,969	(all) 1,949,593	8,523,790	986,473
GPMDB	all	? ^{a)}	?	(all) 8,585,612 (distinct) 491,070 ^{c)}	52,748,666	1,150,085
PeptideAtlas	all	74,600,000	471	(distinct) 40,456	?	285,000
	human	49,000,000	219	(distinct) 12,141	?	97,000
Tranche	all	?	5,502 ^{b)}	?	?	?
GAPP DB	human	130,152	146	(distinct) 2,171 ¹⁾	649,366 ^{f)}	66,655 ^{f)}
Human PP	human	4,567,235	2,695	(distinct) 15,231	1,851,124	?
MAPU 2.0	mouse	?	?	(all) 5,497 ^{d)}	?	?
	human	?	?	(all) 2,926 ^{d)}	?	?
SwedCAD/SwedECD	All	15,897	15,897	n/a	15,897	15,897

In most repositories public data submissions outnumber private ones; in PRIDE, for example, 84% of data submissions were public, and 16% private in December, 2008. In general, submission rates are increasing; see Figure 32 for PRIDE submissions as an example.

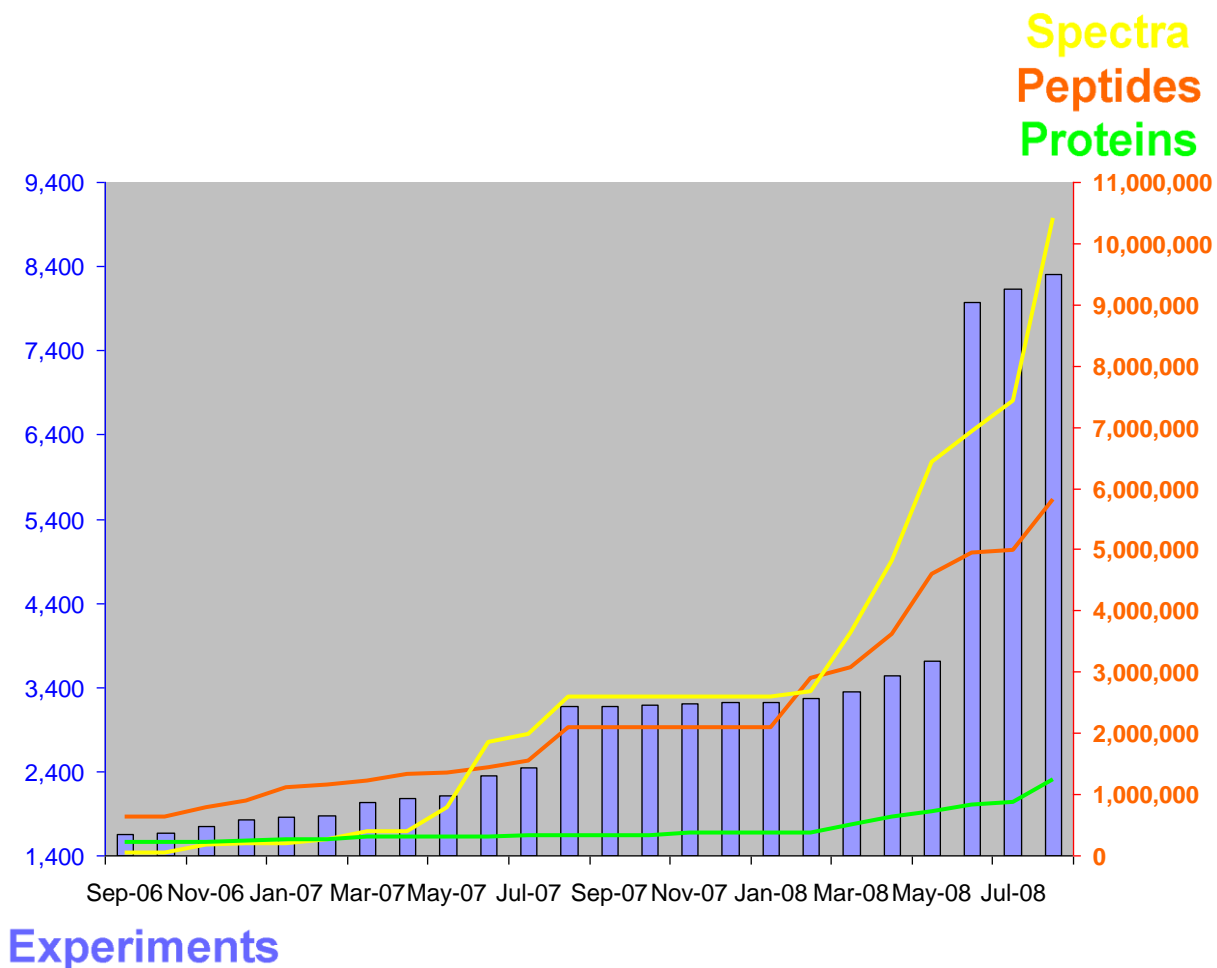


Figure 32 Increase in data submissions to PRIDE repository (Source: Phil Jones ‘Proteomics Standards Development: Progress & Tools’ Lecture, 4th December, Cranfield University)

This increase has most probably been stimulated by the recommendations from journal editors that, upon publication, proteomic MS data should be made public (Table 16).

Table 16 Journals that recommend deposition of MS-based proteomics data into public repositories (autumn, 2008)

Journal	Requirement for submission to public repositories
Journal of Proteome Research Molecular Cellular Proteomics	Authors encouraged to provide access to raw MS data using group websites and public repositories
Nature Methods	Strongly recommends deposition of data before manuscript submission. PRIDE and HPP are mentioned (Editorial, 2008b)
Nature Biotechnology	Recommends proteomics data be posted in public repository before manuscript submission. PRIDE is preferred (Editorial, 2007)
Proceedings of the National Academy of Sciences	Proteomics data is required to be submitted to a publicly accessible database and accession numbers must be provided. Access must be available at the time of publication.
Proteomics	States protein identification results, expression data and MS peak lists should be deposited in a public database. Gives PRIDE as an example.
Rapid Communications in MS	Encourages public dissemination of raw files supporting identifications (Taylor and Goodlett, 2005)

3.7 Standalone versions of some public repositories are available

Some pipelines and their accompanying databases can be installed as standalone versions for local processing and warehousing of in-house data. However, in contrast to web-based versions, which are relatively easy to use and universally accessible, standalones require computing expertise to set up; for example, GPMDB⁹⁰ is available as a standalone, and a complete local PRIDE system (core and the web interface) is available⁹¹ as creation scripts for Oracle or MySQL.

⁹⁰ ftp://ftp.thegpm.org/projects/GPMDB/current_release

⁹¹ <http://sourceforge.net/projects/pride-proteome/>, requires two components: PRIDE core

PeptideAtlas is based on SBEAMS⁹² which can also be downloaded, but it is complex to install because the core biolink module, interface software and proteomics module must be configured individually, and the bundle does not contain all the required Perl and R libraries. Also, the setup files are written for Sybase (proprietary DB system), so it can be difficult to set up a MySQL version, for example.

3.8 Pipelines feed GPMDB, PeptideAtlas and GAPP DB with identifications

There are two routes of data entry into public repositories: by direct submission of peak lists and identifications by users, or via an analysis pipeline. Pipelines, as defined in Chapter 1, perform multi-stage processing to assign identifications to peak lists, as well as other processing steps in some cases. The resulting identifications and peak lists are stored in a repository ('back-end').

3.8.1 GPM

X!Tandem Spectrum Modeler (Craig and Beavis, 2003) is the heuristic search engine in the GPM pipeline. The algorithm produces theoretical spectra for peptide sequences using known relationships between intensity of mass peaks and amino acids, and then matches these with the unknown experimental peak lists using a dot product. This is a

⁹² <http://www.sbeams.org/download/>

'descriptive' approach, because it is based on mechanistic prediction of how peptides fragment (Sadygov *et al.*, 2004).

X!Tandem performs multiple stages of searching and refinement to ensure efficient matching of mass peak lists to sequences, and to optimise for speed. The first step aims to match the theoretical tryptic peptide and fragment masses to the real MS signal peak lists. Then further iterative steps search for PTMs and point mutations. This way, the search space is decreased to a manageable size for computation, and more peaks are successfully assigned.

Also, a new option to 'use sequence annotations' creates a search file detailing the proteins and their potential modifications (referred to annotations) for searching in a more PTM-specific fashion. The files are created using UniProt and GPMDB as sources of annotations; and human, mouse, rat, chicken and yeast are currently supported.

Furthermore, users can now select 'decoy search' on the form to reduce FPs. Users are not restricted to X!Tandem; there is also a consensus spectral search engine X!Hunter (Craig *et al.*, 2006) and a PTP-based search engine, X!P³ (Craig *et al.*, 2005).

3.8.2 TPP

In TPP, spectra are searched against sequence databases using SEQUEST (Eng *et al.*, 1994), or (the recently added) X!Tandem (Craig and Beavis, 2003). The peptide identifications are converted into probabilities using PeptideProphet (Keller *et al.*, 2005) and protein identifications, also with probabilities, are derived from the PeptideProphet results by applying ProteinProphet (Nesvizhskii *et al.*, 2003a). The recent addition is the next SpectraST (Lam *et al.*, 2007) stage, which offers a second round of searching and scoring using a spectral library, followed by Peptide- and ProteinProphet as before. The addition of this search produces more identifications compared to the original TPP alone, and with a low error rate.

The SpectraST search is against an MS/MS library created by combining high scoring spectra to create a database of consensus reference spectra, each one representing an individual peptide. As in X!Hunter, the consensus reference spectrum is compared to the unknown. Individual searches can be performed via the website, where users enter peak lists and mass tolerance. The resulting match of unknown to reference is shown as a comparison view with the x-axis as a 'mirror' where the consensus spectrum points up and the unknown query peaks are mirrored below the x-axis, facing down. The main drawback with SpectraST is that it is only applicable for the PeptideAtlas builds that have sufficient data to create a library.

3.8.3 GAPP

GAPP (Figure 33) is the most important pipeline for this thesis, because the work presented in Chapter 4 and 5 is based on it.

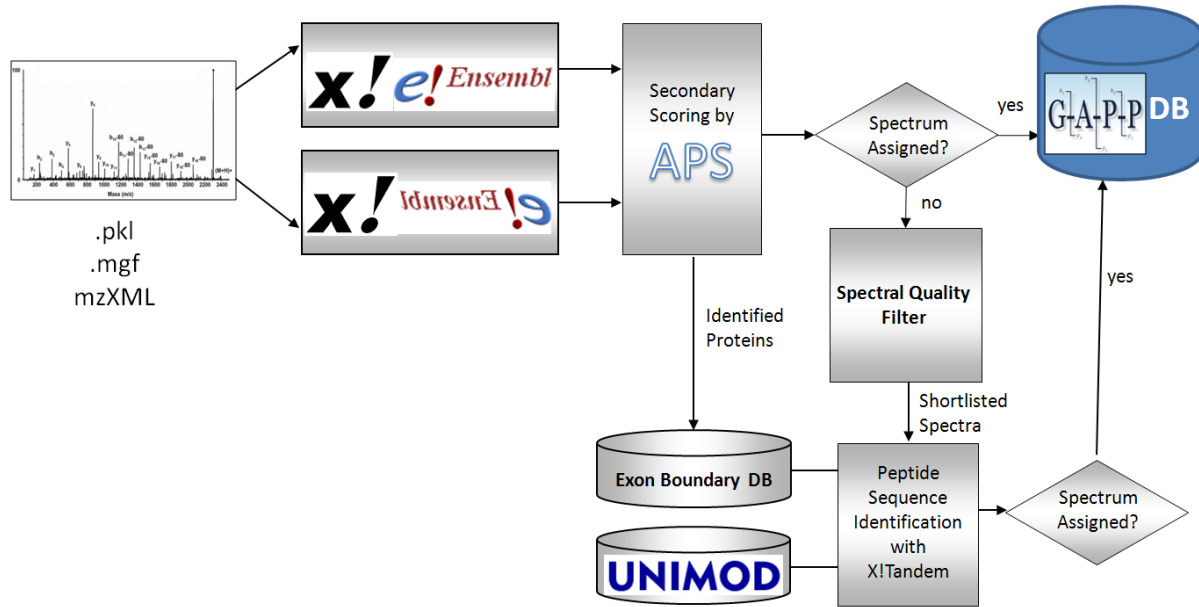


Figure 33 An overview of the Genome Annotating Proteomic Pipeline

As in GPM, GAPP performs peptide scoring using X!Tandem (Craig *et al.*, 2004). Advanced APS filtering (Shadforth *et al.*, 2005b) is then performed to produce high quality protein identifications. GAPP is parallelised, where the process of matching spectra to peptides, using X!Tandem, is split across 16 separate PCs (nodes), which are all connected to a master node with the MySQL database for collating the results from each node. The X!Tandem scores are converted to APS (average peptide scores) values by summing the score for each peptide and dividing by the number of peptides found

for each protein identification. A reversed proteome search is performed (in composite with the target) and the maximum APS found in the reverse decoy is used to filter the target protein identifications found. For a detailed description of APS, see Chapter 4, where GAPP is applied to determine the optimal decoy database design to reduce false positives.

There are splice variant and PTMs loops in GAPP (Figure 33); these aim to increase the proportion of spectra that are successfully assigned to peptides. The pool of unassigned spectra in a given run are filtered by accepting only spectra with 'quality' higher than the worst 'quality' spectrum that was successfully assigned. The quality metric is the sum of the peak intensities divided by the number of peaks, which is usually 100, since only the top 100 peaks are processed by GAPP. The unassigned spectra that pass this criterion are re-searched with X!Tandem against smaller search databases that correspond to the proteins already found in the first round of searching (those in the 'hitlist' table of GAPP DB). The first database is the relevant splice variants for the proteins in the hitlist table (an exon boundary "Pexon" database), and the second is a database of all possible PTM variants (30 of them) for the proteins in the hitlist table. These searches are faster than the original search because the search space is smaller. The Pexon database is created for each target database used in GAPP (human and others) using a Perl script created by Ian Shadforth on a previous EngD project (Shadforth, 2005).

In summer 2008, peptide mass fingerprinting (PMF) functionality was added to GAPP, allowing MS as well as MS/MS data to be analysed, with both results being deposited into GAPPDB.

3.9 Repositories for quantitative proteomics are emerging

Although application of quantitative proteomic MS techniques, such as ICAT, iTRAQ™ and SILAC, is growing, there is still very little quantitative data available in the public domain. This may be attributed to quantitative data not being supported in public systems. Some public offerings are beginning to support quantification. For example, when a complete LC/MS run is uploaded (as a single mzData query file) to ProMEX (Hummel *et al.*, 2007), one of the public pipelines, an entry for each protein/peptide is generated including how often it was identified. The cumulative sum of spectra per peptide and protein may be taken as similar to spectral counting, and spectral counting has been shown to be related to protein abundance, so effectively enables semi-quantitation (Lau *et al.*, 2007). PeptideAtlas also claims to give an approximate estimate of absolute abundance of proteins in biological samples, determined by spectral counting and averaging of the many datasets analysed by its pipeline, the TPP (Deutsch *et al.*, 2008). This information indicates the quantity of surrogate peptide that should be spiked for SRM, for example. Also, after the peptide identification stage of TPP, XPRESS (Han *et al.*, 2001), ASAPRatio (Automated Statistical Analysis on Protein Ratio) (Li *et al.*, 2003) or Libra may be invoked to perform quantitation on suitable data.

However, quantitative data is not available via the public PeptideAtlas repository interface, so is only possible when TPP is installed and run locally.

A major hindrance to the development of repositories to support quantitative data is the lack of formats for its capture, storage and exchange and the variety of strategies available (Lau *et al.*, 2007). Furthermore, best practice in protein quantification has not yet been established, compounding the problem. However, in-roads are being made with iTRAQ reporter ion ratios forming part of the extended PRIDE XML schema (Siepen *et al.*, 2007), and mzML format is expected to support quantitative data, although is unlikely to do so in the first releases. Encouragingly, a recent addition is Quantitative proteomics repository (QuPE)⁹³, a database and algorithmic framework implemented in Java and the Spring framework. It stores proteomics data and metadata from assorted quantitative approaches, such as SILAC, in a consistent fashion, also offering tools for statistical analysis. The resource is accessed via the web interface (after login), which is implemented using Echo2 web framework and has an Ajax-based rendering for graphics.

3.10 Discussion: proteomics data quantity, quality and usage

In summary, the outlook for public proteomics repositories is positive. They continue to grow in content and functionality. However, there is still no complete collection of

⁹³ At http://www.cebitec.uni-bielefeld.de/groups/brf/software/prose_info/index.html. QuPE was previously called 'ProSE'

all publicly available data in one place, despite the ProteomExchange Consortium (including PeptideAtlas, Tranche, GPMDB and PRIDE) being set up to work towards this goal (Hermjakob and Apweiler, 2006).

In general, repositories with the largest amount of data are most useful, because large quantities of data are necessary for meaningful data-mining, deriving consensus libraries for searching, and to improve significance of conclusions derived from the data. However, large file sizes can themselves present a problem for the pipeline-based repositories; GAPP, for example, has an upload limit of ~200MB per submission, although files are routinely much larger, given increased sampling frequency with new MS protocols (to increase sensitivity). Moreover, files are especially large after merging multiple runs for the same experiment. As a result, establishing the best way to submit data is a formidable challenge for the pipeline-based repositories.

In addition, the quality of submissions is an issue for public systems. For spectral quality, research into new methods to assess quality is ongoing (Nesvizhskii *et al.*, 2006), but at present no repository claims to limit public submissions based on any measure of quality. In fact, the move to consensus spectral searches (such as using SpectraST in TPP and X!Hunter at GPM) aims to overcome this issue empirically by averaging across many submissions to account for noise and variability amongst individual spectra assigned to the same peptide.

It is hoped that proteomics repositories may be useful for answering challenging questions such as: Who has observed a set of proteins similar to the set I have observed?, or Can I perform global comparative proteomics using their datasets? To answer the former, the PRIDE Venn diagrams may be used, for example, or GAPP's differential view, or indeed the iSPIDER⁹⁴ integration service (Siepen *et al.*, 2008), which enables users to search and view the identifications made by other groups that are stored in different databases in an internal format called spidyXML. The combined results are then displayed using software clients and specialist viewers.

To answer the latter, global comparative studies are, by definition, only possible if whole proteomes are made publicly available. Examples exist of this scenario: such as Martiens *et al.* who compared protein identifications from brain proteome project (BPP) and three other proteome studies (plasma proteome project (PPP), human platelet proteome and the mouse proteome) (Martens *et al.*, 2006). For now, however, the heterogeneity of data submitted by various labs hampers the process of whole proteome comparison, since sample type, MS instrumentation, identification algorithm and search database can vary. With efforts, such as CPTAC (Blow, 2008) and the Fixing Proteomics Campaign, addressing the issues of reproducibility of data, it is hoped that comparative proteomics will become feasible via public repositories soon.

⁹⁴ <http://www.ispider.manchester.ac.uk/cgi-bin/ProteomicSearch.pl>

Finally, it is likely that in the mid- to long-term future new functionality will lead to better integration between proteomics and other 'omics' data, such as interactomics and metabolomics. This could deliver holistic understanding of biology and the workings of the cell, with implications for improved approaches to biomarker discovery.

3.11 Conclusion

This review described two main types of repository: the analysis pipeline-based repositories (GPMDB, PeptideAtlas and GAPP), and the data warehouse repositories (e.g. PRIDE, Tranche and HPP). It represents novel research that was adapted to form two review articles that were published by Proteomics journal (see Appendix III).

Optimising the design of decoy
search databases using the Genome
Annotating Proteomic Pipeline (GAPP)

4.1 Summary

False positive (FP) identifications can be costly for high-throughput proteomics as they can lead to futile follow-up studies, for example, or erroneous conclusions about the underlying biology. Indeed, inaccurate identification performance provides a reason for investors to continue to be suspicious of high-throughput proteomics, and can lead to problems for those developing new tools to mine the data, once it is stored in proteomics repositories (as is the case for the MRMAid tool developed in Chapter 5).

To remove FPs decoy database searches are routinely applied, the decoy acting as a null model to test if the score of the peptide match is true. Threshold scores, for filtering out FP identifications, are usually set to the highest score achieved by the decoy database search in a given MS/MS data analysis run. Various methods have been published for generating decoy databases, but there is debate about which decoy design is 'the best'. This chapter addresses this question by performing an evaluation of nine diverse decoy designs using public MS/MS datasets from samples of known composition and GAPP pipeline.

4.2 Introduction

Proteomics pipelines are automated, high-throughput workflows that extract protein identifications from m/z peak lists, usually by tandem MS database searching. As already described, one way to assess if the identifications are correct is to apply a decoy database search to filter out false positive (FP) identifications from the target database search results; for example, by taking the highest score achieved by the decoy to siphon off the low scoring peptides in the target search results. Using reversed sequence searches for this purpose is well documented (Peng *et al.*, 2003, Cargile *et al.*, 2004, Qian *et al.*, 2005, Kapp *et al.*, 2005). However, a reverse decoy database may be simple to generate (literally reverse the sequences end to end), but it may not be the optimal choice as regards to false positive rate (FPR) it can achieve; in fact, “*there is no clear consensus ... as to which method for generating a decoy database is best*” (Käll *et al.*, 2008).

To answer the question of which decoy database is best for reducing FPRs, therefore, nine different decoy database designs were systematically investigated, searched in both composite and in parallel to the target proteome. The peptide and protein identification performance was examined for each decoy using Cranfield University’s GAPP pipeline. Unlike other studies to investigate decoy designs (Higdon *et al.*, 2005, Elias and Gygi, 2007, Käll *et al.*, 2008, Reidegeld *et al.*, 2008), the data used in this investigation is a standard protein mixture analysed by multiple laboratories, each of which processed the sample using MS without knowing its composition. It therefore

represents a real-life scenario, taking into account inevitable variability between different experimental setups and researcher experience.

4.3 Method

To perform a decoy optimisation study, a peptide identification pipeline is required so that many searches can be performed in an acceptable period of time. The methodology applied in this study is summarised in Figure 34, and is explained in more detail in the following sections.

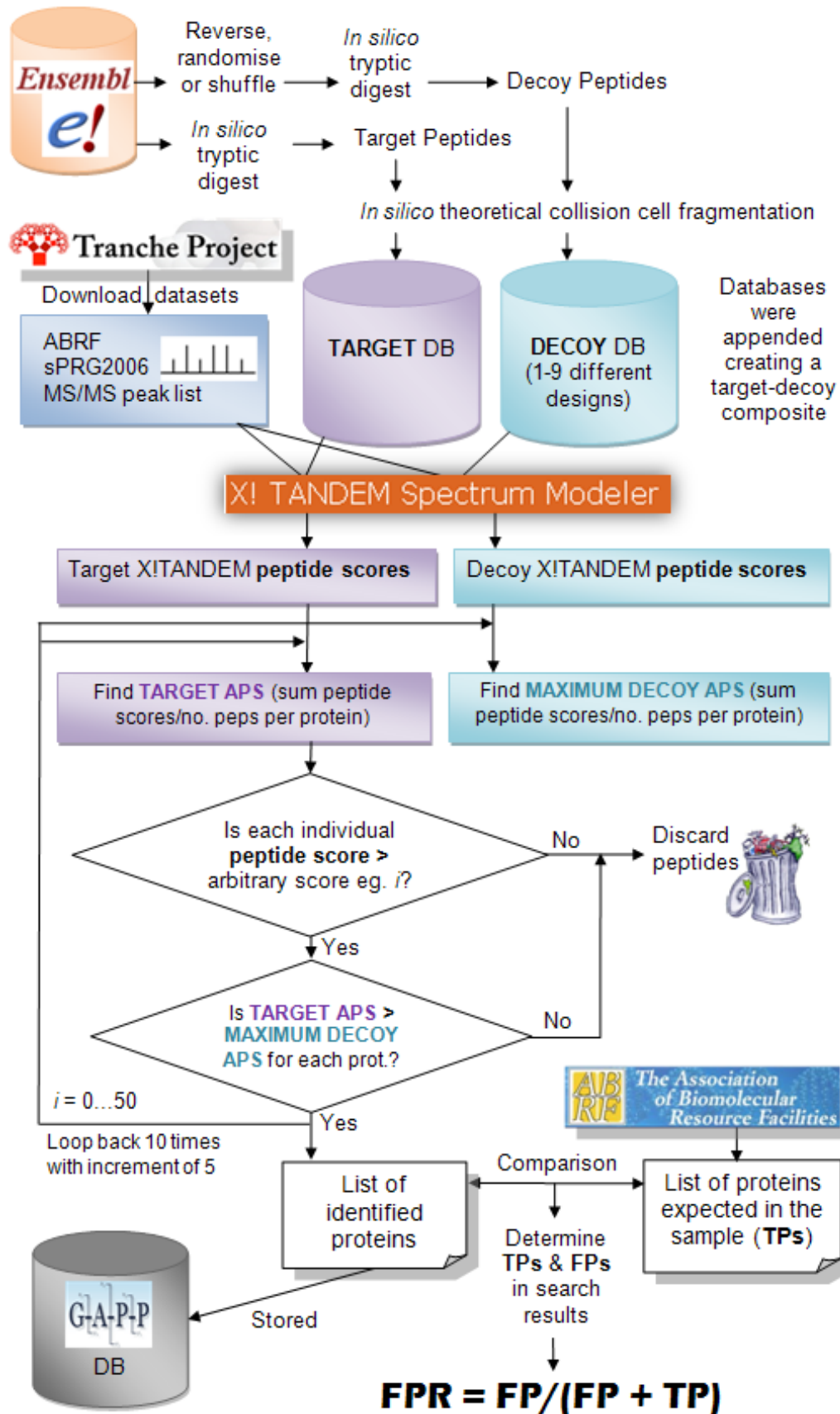


Figure 34 Overview of the approach used to investigate the best decoy database design performed in this chapter. The GAPP-APS pipeline was applied to ABRF standard MS/MS datasets, and performance of the decoy was measured as FPR

4.3.1 Standard datasets with suitable metadata were selected

A prerequisite for the standard data analysed in this study was the availability of accompanying metadata, such as details of the protein content of the samples, chemical treatment, MS instrumentation and the number of spectra:

- Chemical treatment information relates to search parameters, such as variable modifications to include, enzyme regular expression and others
- Instrument type indicates the necessary mass tolerances for the search.
- The number of spectra provided confirmation that all intended spectra from each lab had been downloaded in completeness

The two main contenders, based on fulfilment of these criteria, were the Aurum dataset of 246 human proteins (Falkner *et al.*, 2007) and the ABRF sPRG2006 49 human protein mixture (Andrews *et al.*, 2006); for a full summary of the publicly available datasets at the time of writing go to Appendix IV. The ABRF datasets were chosen because multiple labs had submitted data, so varying levels of data quality could be considered, such as levels of contamination and sample handling. Capturing such variance in data quality was an important factor to include in this investigation, so decoys could be evaluated regardless of individual technique, data quality or mass resolution.

Although the ABRF protein mix was originally intended to contain 49 proteins only, ABRF Proteomics Standards Research Group Bioinformatics Committee (sPRG BIC) confirmed the presence of further ‘bonus’ proteins in the mix (Lane *et al.*, 2007). Subsequently, a ‘master list’⁹⁵ of SWISSPROT confirmed protein constituents was published as a result of retrospective consensus analysis using diverse search strategies. For this thesis, all human proteins on the list were considered, including synonymous accession numbers, where necessary. The total number of proteins was therefore 49 plus associated synonyms totalling 64, plus 40 ‘bonus’ proteins (total 104). Identifications of all these proteins were deemed to be true positives (TPs). To perform the study, the author converted the accession numbers provided by the ABRF to Ensembl-compatible IDs (Appendix IV shows the list).

The ABRF datasets were downloaded using the Tranche Java Downloader Application at ProteomeCommons⁹⁶, and the metadata downloaded from www.abrf.org. Sean Seymour (of ABRF) was contacted to fill in the gaps in the metadata, where necessary. If individual submitting laboratories provided multiple data files for the analysis of the protein mixture, then these files were merged into a single file and the number of spectra was counted and verified against the ABRF metadata file. ReAdW, a free program available for download as part of the TPP software tools, and XCALIBUR (Thermo Scientific, Waltham, MA) were used to convert .raw data to .mgf format,

⁹⁵ Available at www.abrf.org

⁹⁶ www.proteomecommons.org

where required. The ten labs fulfilling the above format and metadata conditions were downloaded; their code numbers were: 00700, 10085, 12874, 14997, 17017, 22069, 25636, 53178, 53908 and 72079.

4.3.2 GAPP-APS pipeline produces high quality protein identifications

The peptide and protein identification performance of diverse decoy database designs was evaluated using the GAPP pipeline (Shadforth *et al.*, 2006), where (as mentioned in Chapter 3) primary scoring is performed by X!Tandem (Craig *et al.*, 2004) followed by validation and protein inference using advanced APS (Shadforth *et al.*, 2005b) (Figure 34).

For the explanation that follows, note that: an analysis 'run' for GAPP is defined as a single MS/MS data submission, with a specific decoy search and a specific set of search parameters; a protein identification (or 'hit') is any protein found in the target database that passes the APS threshold set by the decoy database search for that given analysis run; and a peptide 'hit' is any peptide that was found in the target search database that corresponds to a protein that passed APS.

X!Tandem first assigns a peptide sequence to an MS/MS spectrum in GAPP; the results are peptide sequences each with an X!Tandem score. Proteins are inferred from these peptides, and validated with the advanced APS method. There are two parts to the

advanced APS method, these are: calculating APS thresholds using a decoy database search, and sampling to determine a local maximum for selecting the best APS threshold to use (as shown in Figure 34).

APS threshold and applying decoy database searches to GAPP

GAPP calculates the sum of X!Tandem scores of the peptides that match to the protein, and divides this value by the number of peptides that match to the protein, thus taking the mean average. This aims to account for the fact that the sum of many low scoring incorrect peptide identifications would result in an overall high total score for the protein overall, thus the protein would appear to be a correct identification when it was not; a deleterious situation for an automated search. Thus, peptide assignments to proteins with an APS score below an APS threshold are discarded. This threshold is derived *ab initio* by setting it to the maximum APS score from the decoy database search for the data in question.

GAPP pipeline's default decoy is a reversed version of the target. To create this decoy database, all the amino acid sequences in the Ensembl protein database were literally reversed. The assumption is that the reverse DB covers enough of the possible random sequences to be reliable, which is a credible assumption, but remains to be proven. Despite the uncertainty however, this APS approach in GAPP has been demonstrated to perform well when compared to Sequest, Mascot, PeptideProphet, X!Tandem and others (Shadforth *et al.*, 2006).

Threshold sampling in advanced APS

' i ' is a quality filtering value applied to determine which APS score is best to apply as the filtering threshold. The APS values for each run are computed several times at different increments of i , and the APS value for i that results in the most identifications is chosen. Note that APS is the protein level measure, and i is at the peptide level. For example, the algorithm proceeds as follows:

Step 1: For $i = 0$, if X!Tandem score for the peptide $> i$, calculate the protein's APS score, store the number of peptides and proteins.

Step 2: For $i = 5$, if X!Tandem score for the peptide $> i$, calculate the protein's APS score, store the number of peptides and proteins...etc.

This is repeated in steps for i in increments of 5, up to 50. The final result is a table of values (Table 17) for each value of i , which is stored in the GAPP DB schema.

Table 17 APS_score_info table from the GAPP database schema

Column	Description	Example data entry
unique_id	Unique number of the run, auto-incremented for each data submission to GAPP pipeline	1
single_pass	Number of protein identifications that pass the APS threshold with a single peptide assigned	5
multi_pass	Number of protein identifications that pass the APS threshold with multiple peptide assignments	4
total_pass	Sum of single_pass and multi_pass values. The total number of protein identifications that the pass APS threshold.	9
single_thresh	The APS score required for proteins with a single peptide identification (equal to the maximum APS found for proteins matched to a single peptide in the decoy database search)	76.21
multi_thresh	The APS score required for proteins with multiple peptide identifications (equal to the maximum APS found for proteins matched to multiple peptides in the decoy database search)	6.96
qual_filter	<i>i</i> value (incremented in steps of 5)	10

The APS threshold that is applied for a given run is the one that maximises the total number of identifications: the highest 'total_pass' (Table 17). However, if there are two values of *i* with the maximum number of protein identifications for the same run, then the value of *i* that has a higher number of multiple-peptide protein identifications is applied.

In summary, *i* serves to remove noise (low scoring peptides) in a parametric way by sampling the best APS threshold empirically. This approach increases confidence without losing true positive identifications by finding a local maximum for each search.

GAPP processing capabilities and X!Tandem search parameters

For the decoy database analysis in this chapter, processing was carried out on a Linux-based Beowulf cluster of 16 3.2GHz dual core nodes. In total 1,440 individual search runs were performed, excluding initial testing. The author estimates that this would have taken approximately 90 days to compute using a single processor. Computing power is an important limiting factor in terms of the number of different combinations of algorithms, search parameters and data that can be considered when comparing decoys in this way. For this study, the author collated the spectra to submit to GAPP. Luca Bianco created a script to automate the submission process so the pipeline could be run overnight, taking the spectra from a specified file directory.

4.3.3 Search parameters were applied according to the ABRF metadata

All datasets were processed with the following X!Tandem search parameters: missed cleavages up to 2, carbamidomethyl as a variable modification (because the metadata stated IAA⁹⁷ treatment had been used across all labs) and charge state up to 3+. The mass tolerances were set to reflect the MS instrument used by the individual submitter (Table 18), with five search parameter sets in total. Unfortunately, as for the majority of publicly available MS/MS datasets, the mass accuracies derived from calibration were not reported in the accompanying metadata, so the parameter sets were chosen to

⁹⁷ Iodoacetamide (IAA) is an alkylating sulphhydryl reagent that is used to prepare peptides for MS. It prevents disulphide bonds forming between cysteines. It produces peptides with carbamidomethyl modifications, so the mass shift of these must be accounted for in automated searches.

represent the approximate resolution known to be achieved by the given instrument type (Table 18).

Table 18 The mass tolerance parameters used for each dataset. The tolerances reflect the average resolution achieved by the instrument setup

Parameter set no.	Instrument type	Mass tolerance (Da)	Fragment tolerance (Da)	ABRF submitter code
1	HCT	1.5	0.8	22069
2	LCQ	2	1	72079
3	LTQ	1.5	0.8	00700
4	LTQ-FT	0.3	0.7	12874
				17017
				25636
				53908
5	QTOF	0.5	0.5	10085
				14997
				53178

Ensembl (31st May 2007, homo_sapiens.ncbi36.45.pep.all.fa), which contains all the proteins known to be in the sPRG2006 mixture, was used as the target sequence database for generating the decoy databases.

4.3.4 Two search strategies were applied

The claim of superiority of the target-decoy composite strategy is ubiquitous in the literature (Higdon *et al.*, 2005, Haas *et al.*, 2006, Reidegeld *et al.*, 2006, Klammer and MacCoss, 2006, Balgley *et al.*, 2007, Falkner *et al.*, 2007, Elias and Gygi, 2007, Reidegeld *et al.*, 2008). Despite this, the search method still remains a controversial topic, since separate searches are believed to result in more conservative FPRs, because there is no competition with the high quality matches for the best scores. As a result, decoy hits

may receive elevated scores, so matches to the target have to achieve even greater scores to pass the higher threshold (Elias and Gygi, 2007).

This study performed searches in both ways: firstly, by searching the target and decoy database in parallel so that each database was searched independently from the other, and secondly, using the target and decoy as a composite with the forward search by appending an additional identifier to the decoy entries. The Perl script to append this tag to each accession number was implemented by Luca Bianco.

4.3.5 Nine decoy database designs were investigated

Nine different decoy database designs were tested, four of which were generated at the protein level using the most popular published techniques and five, not previously described, which were produced at the peptide level (Figure 35).

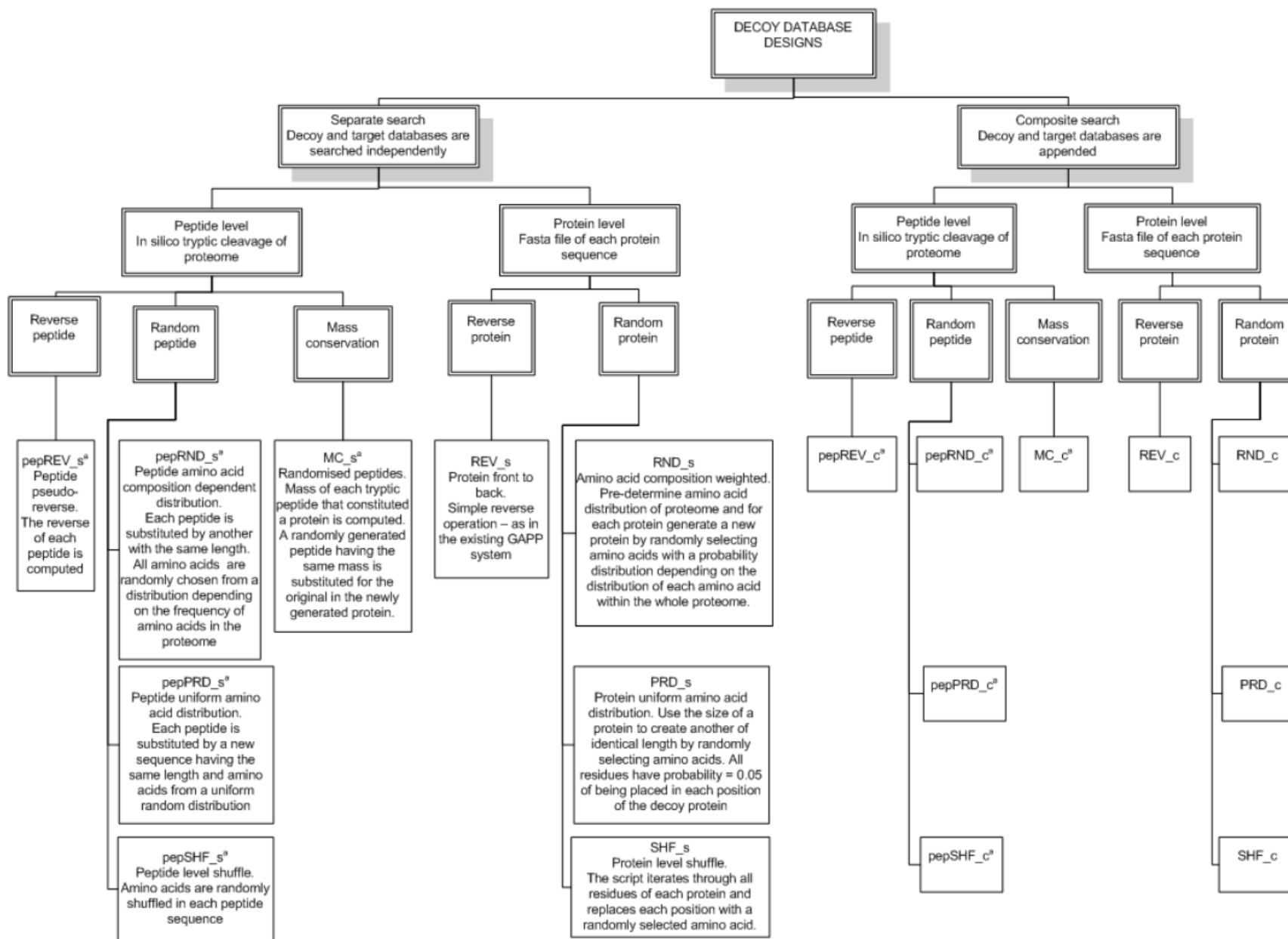


Figure 35 Taxonomy of decoy databases (a) Denotes novel methods of decoy generation implemented in this study. REV: reversed proteome; PRD: uniformly distributed amino acid composition decoy; RND: weighted amino acid composition decoy; SHF: shuffle decoy database; MC: decoy where the mass of peptides are conserved; pep: denotes a peptide level version of the decoy database as described above; ‘_c’ denotes decoys appended to the target proteome; ‘_s’ denotes decoys that are searched separately in parallel to the target proteome.

The main decoy database designs that have been described previously include reversed proteome (Moore *et al.*, 2002, Peng *et al.*, 2003, Cargile *et al.*, 2004, Kapp *et al.*, 2005, Qian *et al.*, 2005, Shadforth *et al.*, 2005b, Reidegeld *et al.*, 2006, Balgley *et al.*, 2007, Reidegeld *et al.*, 2008), shuffled sequence decoys (Stephan *et al.*, 2006, Reidegeld *et al.*, 2006, Klammer and MacCoss, 2006, Reidegeld *et al.*, 2008) and randomised decoys (Elias and Gygi, 2007, Reidegeld *et al.*, 2008). Reverse database searches are the most popular decoy employed to reduce FPs, primarily because they do not require a distribution to be determined, since the reverse proteome is deterministically computed. By reversing each protein, amino acid composition and protein length are preserved, so the statistical properties of the target are maintained.

Shuffle-type operations can also be performed to create decoy databases, where a shuffle is essentially a randomisation process performed within given constraints or rules. The operation may be performed in several ways, for example, where all amino acids of the original protein have been shuffled to random positions using functionality of proprietary decoy database builder software, such as DecoyDB (Reidegeld *et al.*, 2006, Reidegeld *et al.*, 2008). The final alternative is randomising sequences using an amino acid distribution reflecting the entire target proteome.

The properties of the decoy databases employed in this study are summarised in Table 19. For each decoy, excluding reverse sequence decoys, ten instances were generated so that the average distribution of identifications could be calculated. This was required because the randomisation process produced differences in each decoy's composition each time the program to generate it was executed.

Table 19 Summary of the database properties of the decoys. a - note that proteome length is preserved when protein (or peptide) length is preserved. b - in the case of (K/R)P this may be shuffled to create additional K/R not followed by P, which is a tryptic cleavage site.

Decoy design	Protein and peptide length preserved ^a	Amino acid composition preserved in each protein	Amino acid composition preserved in whole proteome	Tryptic cleavage sites preserved	Protein mass preserved	Tryptic peptide mass preserved
REV	Yes	Yes	Yes	No	Yes	No
PRD	Yes	No	No	No	No	No
RND	Yes	No	Approx.	No	No	No
SHF	Yes	Yes	Yes	No	Yes	No
pREV	Yes	Yes	Yes	Yes but may be more	Yes	Yes but may be more ^b
pPRD	Yes	No	No	Yes but may be more	No	Yes but may be more ^b
pRND	Yes	No	No	Yes but may be more	No	Yes but may be more ^b
pSHF	Yes	Yes	Yes	Yes but may be more ^b	Yes	Yes, unless ^b
MC	No	No	No	Yes, but may be more	Yes	Yes

For the reversed proteome (REV) each protein sequence was reversed, meaning that the N-terminus of a protein became the C-terminus, and vice versa. Cleavage sites were not conserved with this method.

For the uniformly distributed amino acid composition decoy (PRD) the algorithm found the size of a protein then created another of identical length by randomly selecting amino acids. A uniform probability distribution was used to choose the amino acid to add to the polypeptide chain. In this way, all residues had equal probability (1/20) of being placed in each position of the decoy protein. The protein length is the only feature preserved by this method.

For the weighted amino acid composition decoy (RND) a pre-computation step was performed to determine the amino acid distribution of the whole proteome. For each protein to be randomised, the size was used to generate a new protein by randomly selecting amino acids with a probability distribution depending on the distribution of each amino acid within the whole proteome. The [0,1] interval was divided in 20 sub-intervals, whose length depended on the frequency of the distribution of each amino acid in the proteome. A random number from a normal distribution between 0 and 1 was chosen to decide the amino acid to add to the polypeptide chain under construction. This method maintained the protein size and the proteome's amino acid

distribution, but cleavage sites and amino acid distribution at the protein level were not conserved.

In the shuffle decoy database (SHF), the positions of amino acids within a protein were randomly changed. This conserved the protein length as well as the amino acid composition. The swapping of residues' positions could, however, affect cleavage sites, which are not protected by this operation.

The peptide level decoys included pREV, pPRD, pRND, pSHF and MC. To generate these, a pre-computation step was required to digest each protein *in silico* into tryptic peptide sequences. Then, a post-computation step reassembled all digested peptides into proteins. In the post-computation step, care was taken to conserve all cleavage sites. The reverse peptide decoy (pREV) was similar to the REV decoy but was performed at the peptide level, whereby the reverse of each tryptic peptide was computed. This preserved the protein length and the amino acid composition. The number and composition of cleavage sites was at least as large as it was in the original protein.

For the uniformly distributed amino acid composition decoy at the peptide level (pPRD), each peptide was substituted by a new sequence having the same length and amino acids from a uniform random distribution. As for the other peptide level decoys,

the cleavage sites were used to determine peptide lengths prior to randomisation. This approach maintains protein length but not amino acid composition. The number and composition of cleavage sites was at least as large as in the original protein.

In the weighted amino acid composition decoy at the peptide level (pRND) each peptide was substituted by another with the same length. All amino acids were randomly chosen from a distribution depending on the frequency of amino acids in the proteome. This preserved protein length but not the amino acid composition of peptides. The number of cleavage sites was at least as large as in the original protein.

For the peptide shuffle decoy (pSHF) amino acids were randomly shuffled in each peptide sequence using the same approach as for SHF. This preserved protein length and the amino acid composition of peptides. The frequency of cleavage sites was at least as often as in the original protein.

Finally, for the mass-conserved decoy (MC), the mass of each tryptic peptide that constituted a protein was computed and a new, randomly generated, peptide having the same mass (plus/minus a tolerance, in this case fixed to 0.5 Da) was substituted for the original in the newly generated protein. This method preserved the mass of digested peptides but varied the amino acid composition and the protein length.

Palindromic sequences were not removed, neither were any target sequences found in the decoy by chance. This is because it has been shown that palindromic peptides account for a negligible proportion of the proteome (Elias and Gygi, 2007) and the probability of obtaining an identical sequence by chance is small $(1/20)^l$, where l is sequence length.

For this part of the work, the author designed the selected decoy models, and performed conceptual design of the algorithms to create them in conjunction with Luca Bianco, who coded the final programs to generate the decoy database instances: a program written in C for the mass conservation decoy and Perl scripts for the others.

4.3.6 Decoy database performance was measured using FPR

False positive rate (FPR), $FP/(FP + TP)$, was used as the principal measure of decoy performance. This is because it captures the identification performance of each search, taking into account both correct and incorrect identifications. FPR could be determined because a standard protein mixture was analysed (Higdon *et al.*, 2005). FPR was calculated at both the protein and peptide levels, however the focus of the analysis was predominantly on protein identification performance.

There is ambiguity in the community as to the definition and usage of the term 'False Positive Rate' (FPR) versus 'False Discovery Rate' (FDR). In the literature there are instances where FPR and FDR have been used interchangeably, for example in Elias and Gygi's paper (Elias and Gygi, 2007), where FPR was defined as $FP/(FP + TP)$ (as in this thesis) and in (Jones *et al.*, 2008a) where the same measurement is referred to as FDR.

Further complications arise still, for example, where FDR is defined as the percentage of PSMs (peptide-spectrum matches) that are incorrect by some authors, such as (Käll *et al.*, 2008), and can be applied when datasets of unknown composition have been used (Blackler *et al.*, 2006, Reidegeld *et al.*, 2008). Also FDR - for separate target decoy searches - may be calculated as the number of decoy peptides identified, divided by the number of target peptides identified (Choi and Nesvizhskii, 2008a, Tabb, 2008). The purpose of FDR in this case is to give an indication of the percentage of incorrect peptide identifications that have been accepted as correct by passing a user-defined threshold. In this study, however, FDR, by this definition, did not need to be determined, because a standard dataset of known composition was employed, so FPR sufficed - as calculated as $FP/(FP + TP)$, where the TPs list was available.

4.3.7 Statistical analysis included factorial ANOVA

ANOVA was performed on the mean FPR values to determine if there was a statistically significant difference in performance across the decoy designs. It was performed with three factors using Genstat (VSN International Ltd., UK), the factors being: i) decoy design (nine as shown in Table 19), ii) search strategy (composite and separate) and iii) instrument (five MS setups as shown in Table 18). The aim was to determine which, if any, decoy design was significantly better than the others and also to investigate the effect of search strategy and instrument type on FPR - to put any difference between decoys into context. To establish significant differences, pair-wise comparisons of least significant difference (LSD) values were made. For instrument

type, the number of replicates was not uniform so LSDs had to be compared taking into account the replicate number for each case. Bespoke scripts were written by the author to import data into Matlab (Mathworks, Natick, MA) and Genstat for these analyses.

In addition to ANOVA, box whisker plots were generated to graphically illustrate the differences in protein identification performance across different labs and decoys. Such a plot shows the robustness of each decoy design, demonstrating the variability between the ten individual instances. These plots were generated using Matlab scripts written by the author. Additional graphs were created using Excel.

To explore the effect of decoy design on identification performance further, the author designed a database to capture several other metrics at the time of search. These properties were designed to give further, detailed understanding into the aetiology of the differences in decoy design performance, and included:

- Number of protein identifications (TP, FP and decoy database)
- Number of peptide identifications (TP, FP and decoy database)
- Number of peptides found per protein identification (TP, FP and decoy database)
- Number of peptides matched with seven or fewer amino acids (TP, FP and decoy database). Seven amino acids is the number reported by Elias and Gygi at which peptide redundancy between target and reverse decoy deteriorates (Elias and Gygi, 2007).
- Average length of peptide identifications (TP, FP and decoy database)
- Number of assigned peptides identical in sequence across target and decoy
- False positive rate at the peptide level
- False positive rate at the protein level
- APS score threshold for single peptide identifications
- APS score threshold for multiple peptide identifications

For each decoy database (except the reverse decoys) there were ten values for each metric in this list, these represented the ten instances. The author implemented the MySQL database to capture these metrics, and Luca Bianco wrote the intermediary script between GAPP and this metrics database. The author manually queried specific values, such as APS thresholds, from the metrics database schema.

4.4 Results

4.4.1 FPRs summary

Mean FPRs were reported between zero and 5.6 percent for peptide level identifications, and zero and 17.9 percent for the protein level identifications. Actual numbers of FPs and TPs used in the calculations are provided in Table 20. The separate

search strategy generally performed better, producing lower FPRs over the majority of decoy designs (Figure 36 and Table 20).

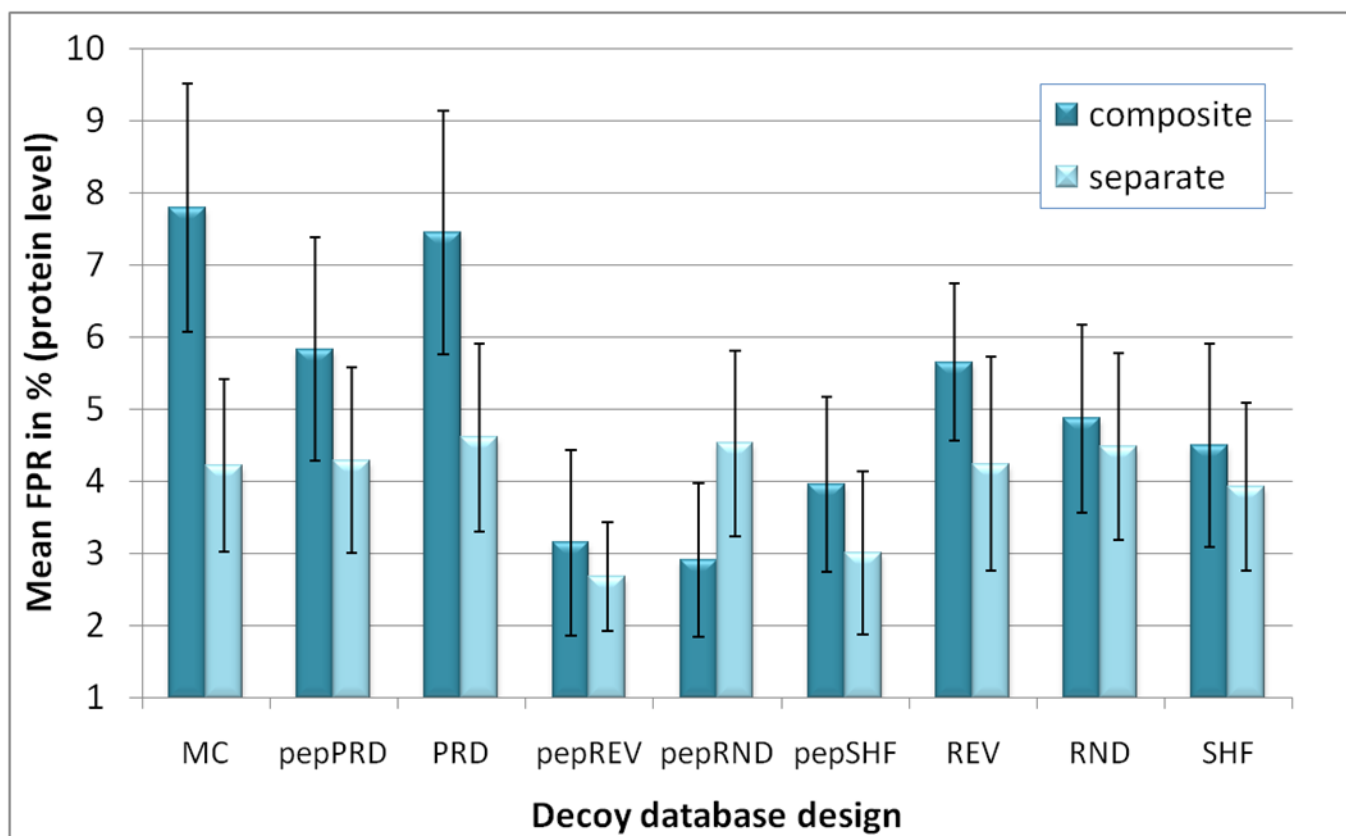


Figure 36 Mean protein identification FPR by decoy design and search strategy. Separate searches are generally more accurate. The majority of decoys produced a lower FPR when searched independently of the target database, the only exception being pRND which resulted in lower FPR when searched in composite with the target. pREV searched in composite with the target appears to be the optimal method, followed closely by pSHF in both composite and separate. Bars represent the standard error. Decoy/method LSD 5% = 2.86

Table 20 Mean true positives and false positives across all data submitters and decoy search types.

(a) TP proteins (grand mean = 32.1, standard deviation = 13.1)

Submitter	c									s									MC_s
	REV_c	PRD_c	RND_c	SHF_c	pREV_c	pPRD_c	pRND_c	pSHF_c	MC_c	REV_s	PRD_s	RND_s	SHF_s	pREV_s	pPRD_s	pRND_s	pSHF_s		
22069	34	34.4	34.8	34.3	33	33.8	33.6	33.6	35.2	34	33.1	33.4	32.8	33	33	33.8	32.9	33.5	
72079	61	64.5	63.3	62.4	62	63.2	62.2	63	64.1	58	63.1	62.1	61.7	62	61.9	62	62.1	61.7	
00700	20	22.7	20.4	19.8	20	21	19.7	19.8	22.5	19	22	22.9	21.1	24	23	23.1	20.9	21.7	
12874	36	35.8	35.2	35.7	36	35.3	35.8	35.9	36	36	35.8	35.7	36	36	35.2	36	35.9	36	
17017	34	33.7	33.3	33.8	32	33.6	33	33.3	33.8	34	33.5	33.6	33.4	34	33.4	33.6	33.6	33.4	
25636	10	9.5	8.4	8.7	7	8.5	7.8	9	8.9	9	10	9.5	8.8	8	9.1	7.8	8.9	8.6	
53908	28	27.1	26.2	26.3	24	26.8	26.1	25.9	27.1	27	26.4	26	26.4	26	26.3	26.7	26.4	26.7	
10085	37	37.2	37	37	37	37	37	37.1	37.1	37	37	36.9	36.8	37	36.8	36.9	37	37	
14997	37	36.4	36.7	36.8	36	36.5	36.6	36.4	36.7	36	36	35.5	36.7	36	36.1	35.9	35.4	36	
53178	25	27.1	24.9	25.5	26	26	25.6	25.1	26.5	26	25.5	25.3	25.8	27	25.8	26.3	25.7	25.9	

(b) FP proteins (grand mean = 1.80, standard deviation = 2.61)

Submitter	c									s									MC_s
	REV_c	PRD_c	RND_c	SHF_c	pREV_c	pPRD_c	pRND_c	pSHF_c	MC_c	REV_s	PRD_s	RND_s	SHF_s	pREV_s	pPRD_s	pRND_s	pSHF_s		
22069	3	3.5	3.2	2.6	0	2.5	1.8	1.7	4.3	3	2	2.4	2.2	0	1.5	2.3	1.5	2.3	
72079	8	14.1	9.4	9.8	3	11.2	7.6	6.9	12.7	2	10.3	9.6	8.5	3	9.7	9.8	8.2	7.8	
00700	1	2.6	1.3	0.2	1	1.6	0.3	0.3	3	0	1.6	2.1	1.3	2	2.2	2.3	1	2.2	
12874	0	0.2	0	0.1	0	0	0	0	0.3	1	0.1	0.3	0.2	1	0.6	0.7	0.1	0.2	
17017	2	4.4	1.8	1.6	2	2.7	1.6	1.2	4.7	6	1.6	1.7	2.2	1	2	2.1	1.7	2.2	
25636	1	1.1	0.8	1	1	1.2	0.3	1.2	1.2	0	0.8	0.4	0	0	0.3	0	0.1	0.1	
53908	1	1.4	0.5	0.6	0	0.6	0.1	0.6	0.9	2	0.6	0.9	0.4	0	0.3	0.5	0	0.5	
10085	1	0.8	0.9	0.6	0	0.6	0.3	0.6	1	0	0.5	0.7	0.8	1	0.5	0.6	0.3	0.7	
14997	3	1	0.8	0.8	0	1	0.2	0.9	1	1	0.4	0.1	0.8	1	0.5	0.5	0.2	0.4	
53178	1	1.4	0.3	0.6	1	0.8	0.4	0.5	1.5	1	0.7	0.3	0.7	1	0.5	1	0.5	0.7	

The grand mean FPR across all 180 means (10 data submitters x 9 decoy designs x 2 search strategies) at the protein level was only 4.56 percent (standard error 0.31), and at the peptide level only 1.34 percent (standard error 0.10).

4.4.2 Recommendation for decoy design based on protein level FPR

The ANOVA F probability values were under 0.05 for each of the three factors for the protein level FPR mean values: 0.030 for decoy design, 0.021 for search strategy and <0.001 for MS instrument setup, respectively. This was an exciting result indicating that there were significant differences between the FPRs across the decoy designs, and that the FPRs obtained by separate searches were significantly different to FPRs obtained by composite searches over the datasets tested. Importantly, the interaction F values showed no statistically significant interactions between the three factors at the protein level, with all values exceeding 0.05, meaning that the decoy effects on FPR were reproducible across the different search methods and instrument types. This is important as it indicates that the relative performance of the decoys across the samples analysed in the study is independent of the MS instrument or search strategy employed. Finally, inspection of the residuals showed that the average FPRs fitted the normal distribution, so the underlying assumptions required for ANOVA were valid.

To determine which decoy design performed best, the LSDs at 5% from the ANOVA analysis were compared to the mean FPR values (Table 21). The LSD at 5% is the absolute value by which two means must differ to be deemed significantly different

with 95% confidence. The comparison shows that pREV, pSHF and pRND were significantly better in terms of protein FPR than MC and PRD. pREV was also significantly superior to REV and pPRD. There was, however, no single 'winner' decoy database that was significantly better than all the rest; however, the results are suitable to make the recommendation that peptide level reverse be used routinely for automated searches using the APS method. This is because pREV is as good as the other decoy designs, significantly better than MC, PRD, pPRD and REV, but, perhaps more importantly, in practical terms it is easier to generate and use than pSHF or pRND (the other contenders) because it does not require multiple database instances to be generated to derive a robust result.

Table 21 Significant differences between decoy designs for protein identification false positive rate (FPR). The underlined values are the pairs that demonstrate statistical significance. To determine significant differences the least significant difference (LSD) value at 5% was compared to the mean FPRs. The LSD at 5% is the absolute value by which two means must differ to be deemed significantly different with 95% confidence. pREV, pSHF and pRND are significantly better in terms of protein FPR than MC and PRD, and SHF is significantly better than MC. pREV is also significantly better than pPRD and REV. It is therefore the recommended decoy design for protein identification because it is as good as the other decoy designs, significantly better than those mentioned and is easier to generate than other randomised decoys, because it does not require multiple instances to be run to derive a reliable result. LSD at 5% = 2.022

Decoy design		MC	pPRD	PRD	pREV	pRND	pSHF	REV	RND	SHF
	Mean FPR	6.01	5.06	6.03	2.91	3.72	3.49	4.95	4.68	4.22
MC	6.01	x	0.95	0.02	<u>3.1</u>	<u>2.29</u>	<u>2.52</u>	1.06	1.33	1.79
pPRD	5.06		x	0.97	<u>2.15</u>	1.34	1.57	0.11	0.38	0.84
PRD	6.03			x	<u>3.12</u>	<u>2.31</u>	<u>2.54</u>	1.08	1.35	1.81
pREV	2.91				x	0.81	0.58	<u>2.04</u>	1.77	1.31
pRND	3.72					x	0.23	1.23	0.96	0.5
pSHF	3.49						x	1.46	1.19	0.73
REV	4.95							x	0.27	0.73
RND	4.68								x	0.46
SHF	4.22									x

The ANOVA provides a Boolean result on significant difference, but to further explore the variability in decoy design the box whisker plot is useful (Figure 37, continued in the Appendix IV). This plot shows that certain decoy generation methods produce more reproducible performance than others. This variation further affirms the recommendation of a reverse decoy design, where standard deviation is always zero.

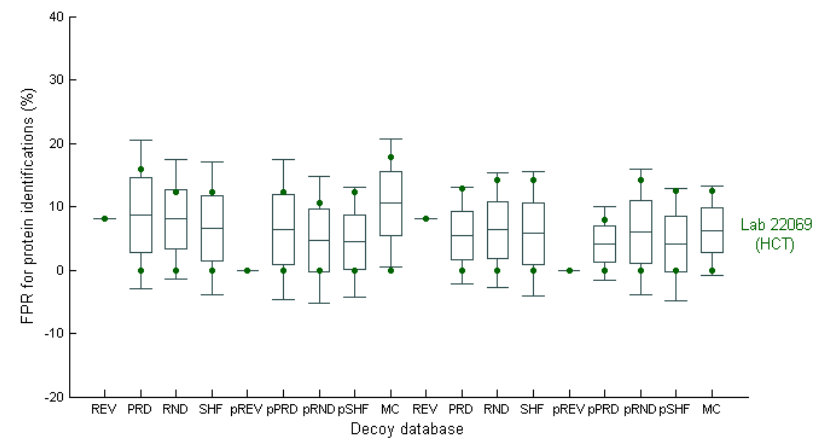
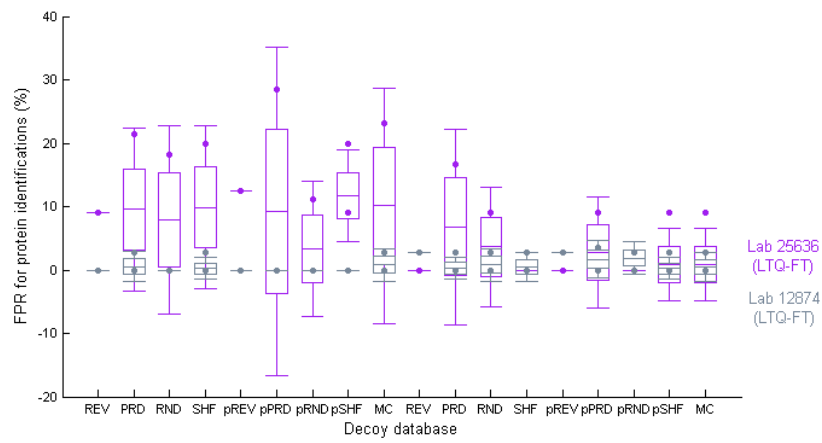
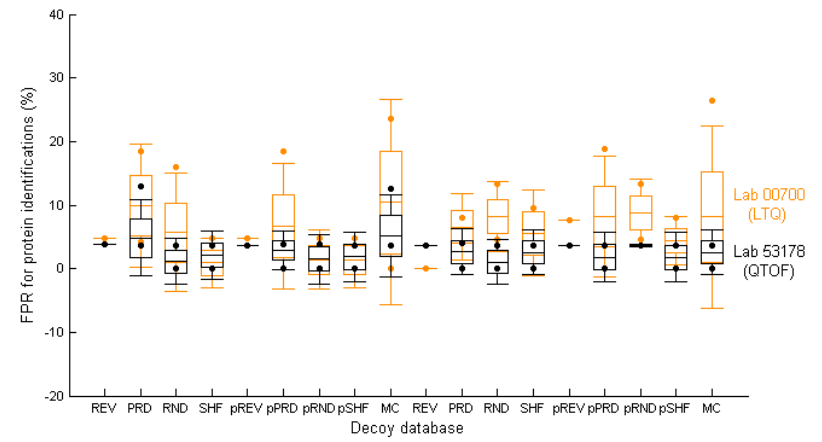
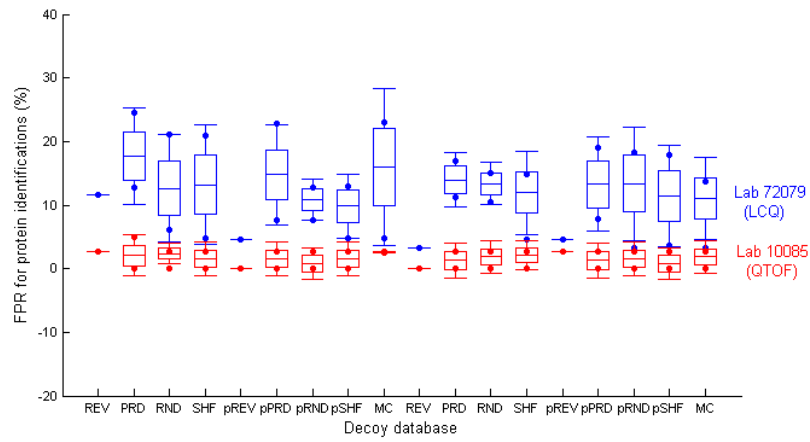


Figure 37 Box whisker plot to illustrate the distribution of protein level false positive rates (FPRs) across the ten database instances for each decoy design. Each color represents an individual ABRF data-submitting laboratory. The horizontal line is the mean of the ten database instances, the box around the line shows one standard deviation above and below the mean, the whiskers show two standard deviations from the mean and the filled dots are the maximum and minimum individual FPR values obtained for the given decoy. The first nine decoys on the x-axis were searched in composite with the target and the last nine separate to the target. A representative of each instrument type is shown, for the remaining three labs (53908, 14997 and 17017) see the Appendix IV.

The coefficient of variation (CV) value for protein FPR means in this study was 70.6%, which is why small differences between decoy designs could not be detected as significant by the ANOVA. This is also the reason for the relatively large LSD values - with FPRs having to be very different to be classed as significantly different (Table 21). However, the data volume and scope was sufficient in this study to make statistically significant, meaningful recommendations, hence sub 0.05 F values were reported.

4.4.3 Recommendation for search strategy based on protein level FPR

For search strategy, the ANOVA generated an LSD (5%) for FPR at the protein level of 0.953. The difference between the composite and separate means, 5.12 and 4.00, respectively, was 1.12, therefore composite and separate search methods were significantly different across the datasets included in this study - with separate being the significantly superior method. The recommendation, therefore, is that separate searches should be used, which is in contrast to the multiple studies using datasets from samples of unknown composition that recommend composite searches, such as (Elias and Gygi, 2007, Reidegeld *et al.*, 2008).

4.4.4 Recommendations based on peptide level FPRs

For peptide level FPRs, F values were not below 0.05 for decoy and search method: 0.077 and 0.178, respectively. Machine type, however, did show the expected significant difference at the peptide level (F value <0.001). Unlike the protein level FPR, the peptide FPR values also showed interactions between search strategy and instrument (F value <0.001). The meaning of this interaction is very difficult to interpret, however more pertinent to this study is the fact that there were no interactions between decoy design and search strategy, and no interactions between decoy design and instrument.

In contrast to the protein level, there was no statistically significant difference between composite and separate search strategies at the peptide level, although the mean overall peptide FPR was lower for the separate search data (1.2) than the composite (1.4). Finally, the CV value for peptide FPR means was 79.5 percent, which is higher than the protein level.

4.5 Discussion

In this study, the FPRs at the peptide level were lower than the protein level. This is to be expected, because the peptide hits are identifications corresponding to actual matches between the database and m/z values, whereas protein identifications are inferred from peptide level data, and the process of inference is subject to additional error and ambiguity.

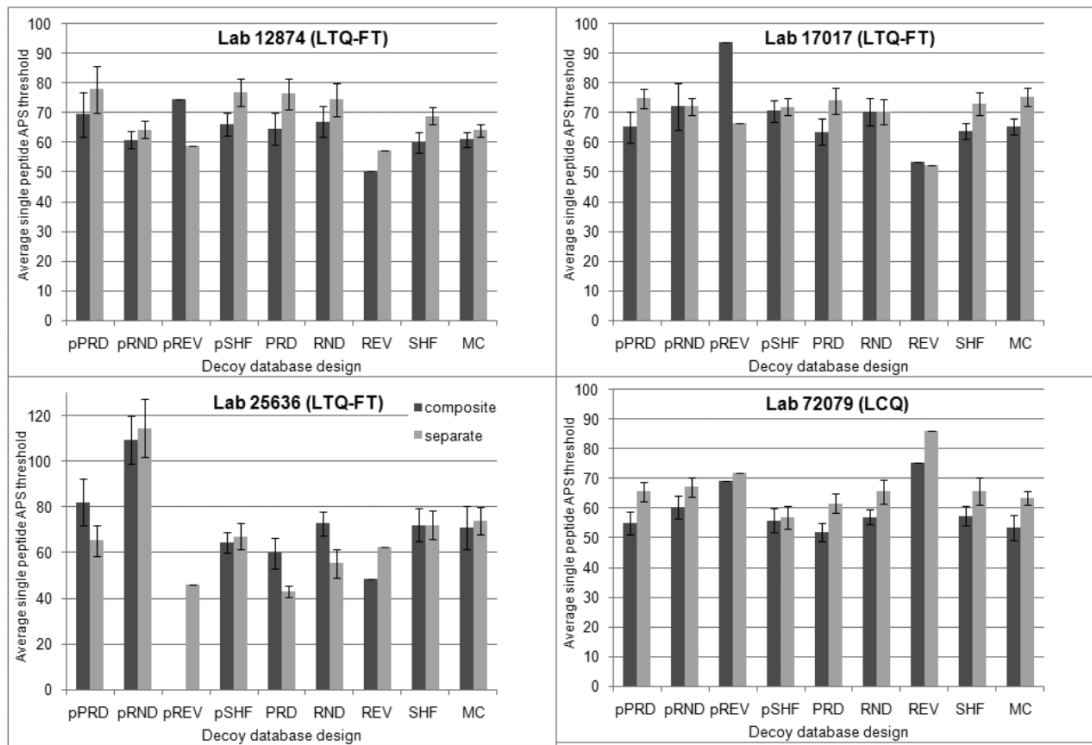
4.5.1 APS threshold explains the differences in FPR between the different decoy database designs

To understand the reasons for the lack of a clear 'winning' decoy design, and to see why pREV, pSHF and pRND were significantly better than MC and PRD, various metrics were examined across the different decoy designs. The APS threshold was the metric that best explained these observations. Figure 38 (a) illustrates average single-peptide APS threshold for four of the ABRF data submitters (across the ten instances of each decoy design), and Figure 38 (b) the same only for multiple peptide APS thresholds. Arguably, the single threshold is the most important threshold to examine because the results showed that the average number of peptides required to make a false positive identification was one, compared to true positives which on average found multiple peptides to assign to the protein (data shown in Appendix IV). It shows that across the four labs shown pREV, pSHF and pRND have some of the highest APS thresholds compared to the other decoys, and explains why these three had significantly better FPRs than MC and PRD (Table 21). MC and PRD routinely had lower APS thresholds compared to pREV, pRND and pSHF, so these

decoys permitted more low-scoring FP identifications to pass, thus producing higher FPRs as a consequence.

Figure 38 also shows that the best decoy was not always the same (in terms of threshold) across the different datasets. Lab 17017, for example, showed pREV had the highest, whereas 25636 showed pRND had the highest for single peptide APS thresholds. This study, however, looked holistically across ten different datasets and the statistical analysis showed no significant interactions meaning that overall the decoys behaved reproducibly across the datasets - despite this sample perhaps indicating otherwise.

(a)



(b)

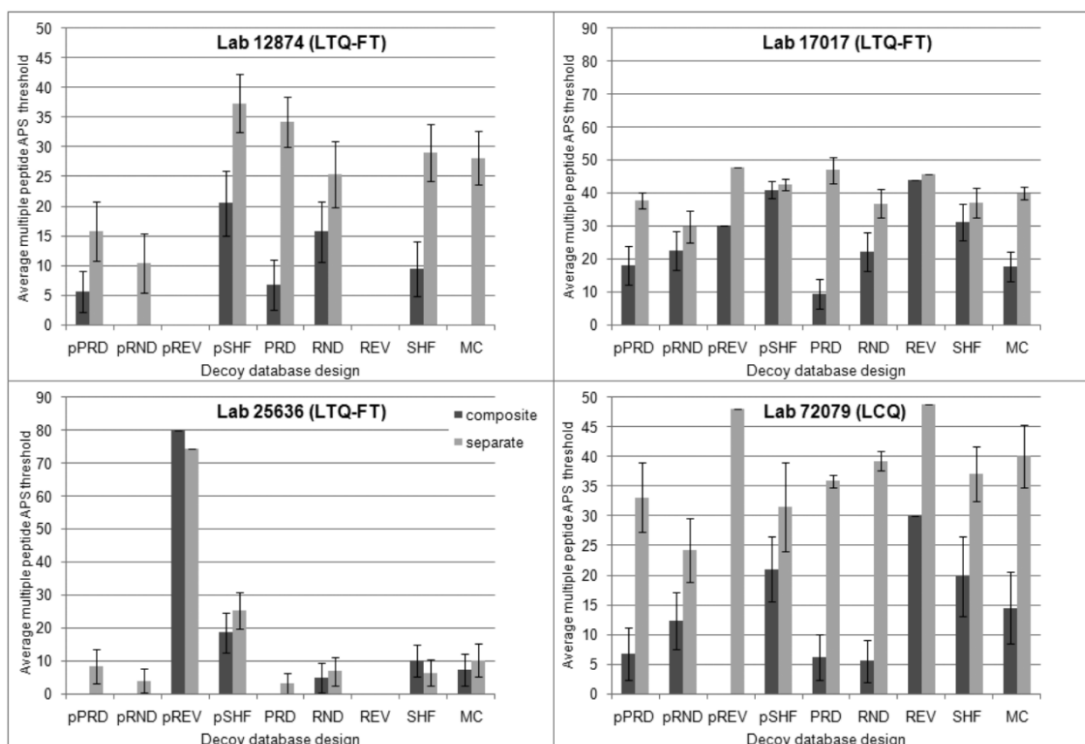
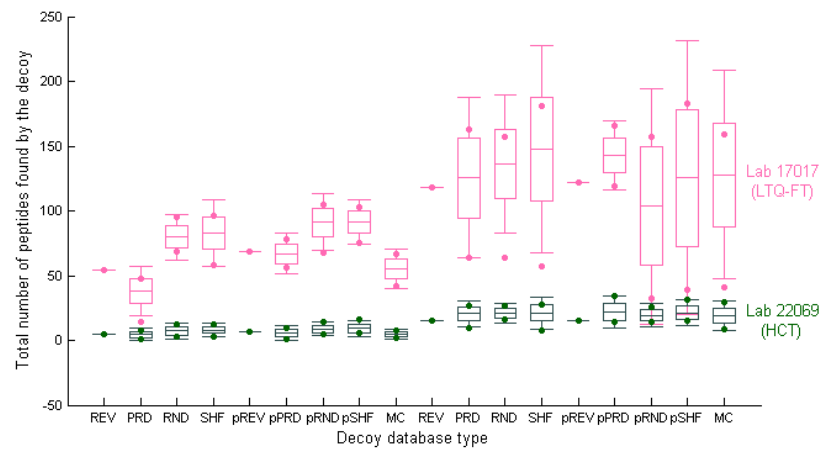
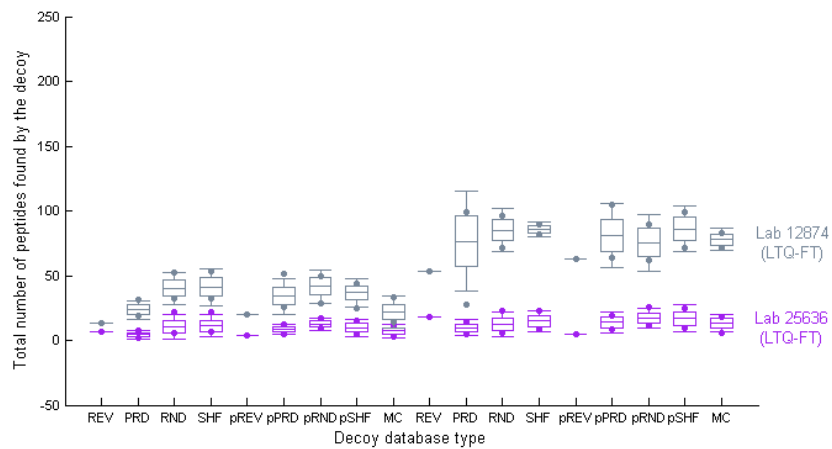
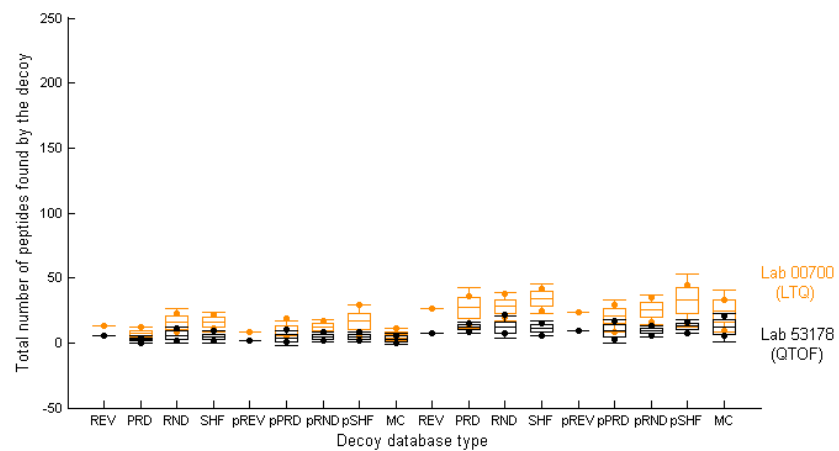
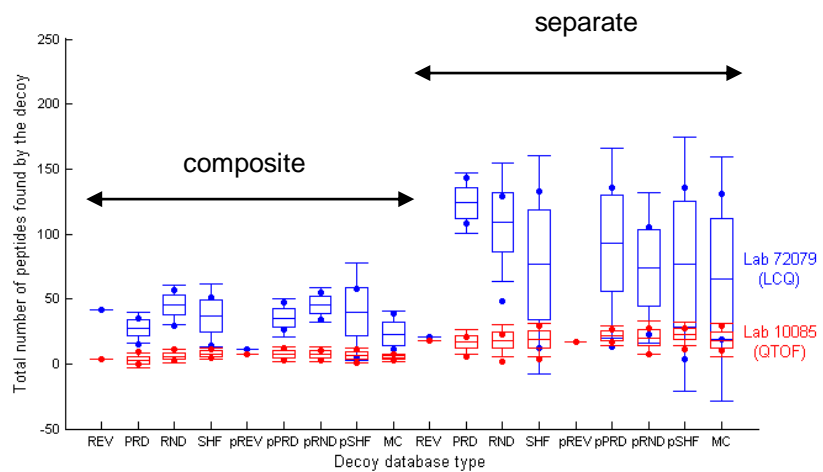


Figure 38 Average APS score thresholds across the diverse decoy designs for (a) single-peptide protein identifications and (b) multiple-peptide protein identifications. Four of the ten ABRF submitting labs are shown and were chosen to represent a suitable range of candidates, including data from different instruments and across different ranges of FPs (72079, high FPs; 12874, low FPs; 17017 and 25636 mid-range FPs). The remaining three labs (00700, 14997 and 22069) are shown in Appendix IV. The bars show the standard error above and below the mean; reverse decoys have zero variation so have zero error. The thresholds are plotted to establish the underlying reasons for the differences in decoy database performance; in particular, to understand why pREV, pRND and pSHF were significantly better than MC and PRD. The figure shows average peptide score (APS) thresholds, where APS is calculated as the sum of the individual peptide X!Tandem scores divided by the number of peptides for the given protein identification. This figure illustrates average single-peptide APS threshold (across the ten instances of each decoy design). The single threshold (in (a)) is particularly important because the results showed that the average number of peptides required to make a FP identification was one or less (remembering there are ten instances of each decoy design some of which had no FPs), compared to true positives which on average found multiple peptides to assign to the protein (see the Appendix IV for these data).

Finally, Figure 38 clearly shows that separate searches achieve higher (thus more stringent) APS thresholds. This is because the decoy hits are not in direct competition with high quality matches when searched independently, so can achieve elevated APS scores (Elias and Gygi, 2007). Lastly, pREV and pSHF conserve both peptide and amino acid composition (see Table 19 earlier), this could be an additional contributory factor for these decoys performing significantly better than MC and PRD.

Other metrics that explained some of the observations were: (1) total number of peptide hits (Figure 39), which showed that more hits are made against the decoy databases when searched in parallel, as opposed to composite; and (2) average length of peptides used in identifications (Figure 40), which showed that on average false identifications are made using shorter peptides compared to those for true identifications. This data also indicates that the decoy databases represented acceptable null models, because peptides of very similar length to TPs were found in the decoys.



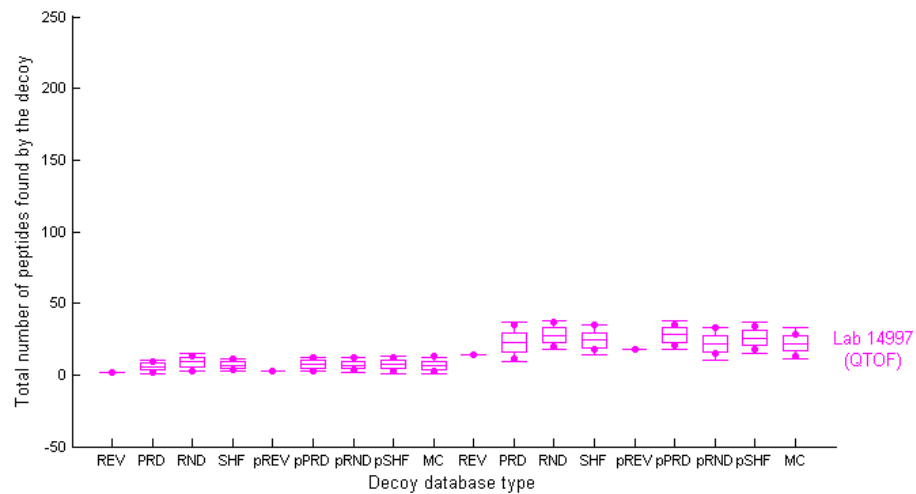
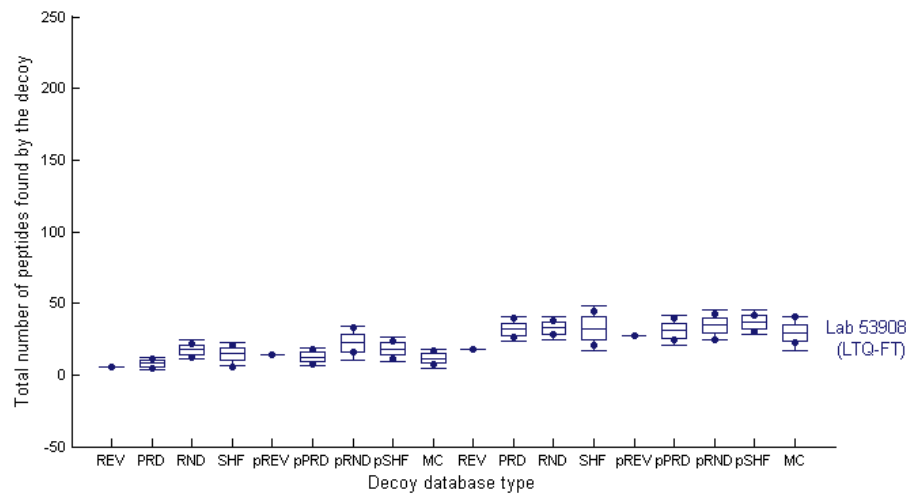


Figure 39 Understanding differences between decoy designs by examining the number of peptides found by each individual decoy design. The horizontal line is the mean of the ten database instances, the box around the line shows one standard deviation above and below the mean, the whiskers show two standard deviations from the mean and the filled dots are the maximum and minimum individual FPR values obtained for the given decoy. The first decoys on the x-axis were searched in composite with the target and the last nine in parallel. More peptides are found in separate searches because there is less competition for matches. This is also why there is a more conservative hit rate with the separate search because with more peptides found by the decoy, there is more likely to be a higher maximum APS (hence higher APS threshold) in the decoy meaning more peptides hits are filtered out. (The remaining two labs are plotted separately because the values overlap).

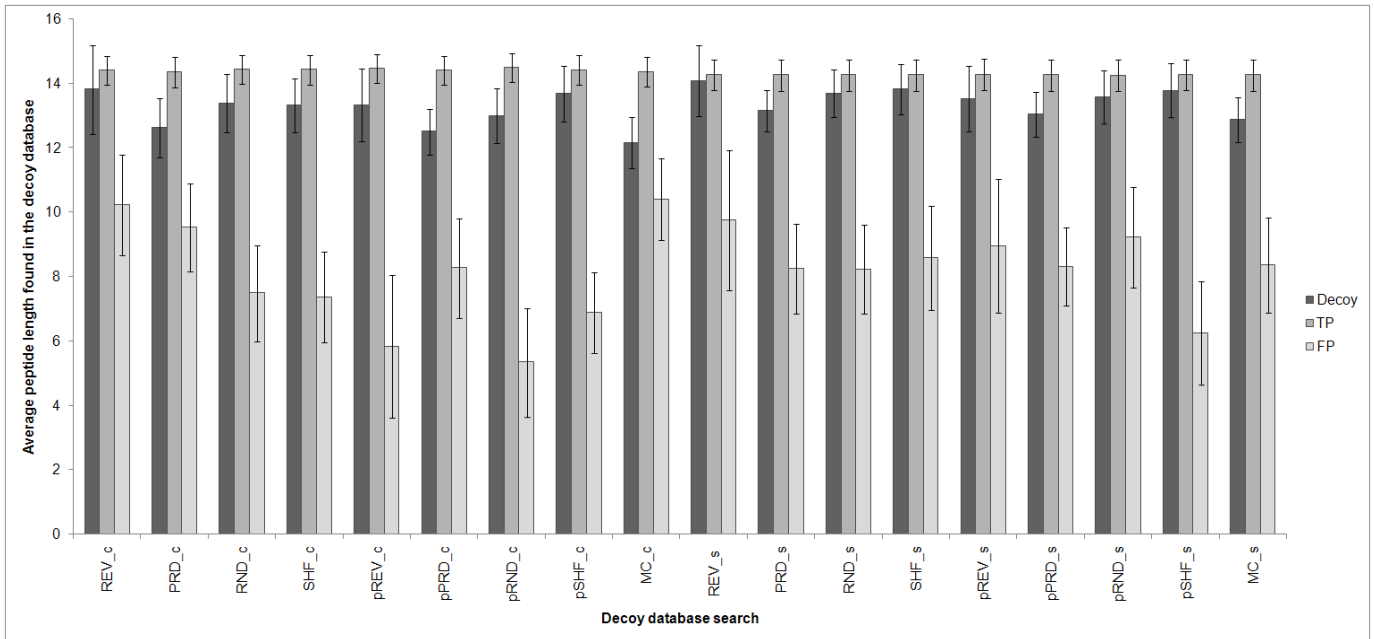


Figure 40 Mean peptide lengths found across ten instances of each decoy type. The bars show the standard error above and below the mean. False positives are generally attributed to shorter peptide matches.

The remaining metrics that were considered included the mean number of assigned peptides that were identical between the target and decoy- this was virtually zero (grand mean 0.018); and the number of assigned peptides of seven amino acids or less, which showed that on average only one or fewer (0.83) peptides under seven residues was hit in the TP protein space and only 0.13 in the FP space across all decoys and labs.

4.5.2 Identification performance was comparable with the original ABRF study

Comparison with the results reported in the original ABRF analysis poster⁹⁸ have to be limited to the 49 core proteins because the ‘bonus’ constituents were only identified subsequently (Lane *et al.*, 2007). The comparison showed that on the whole, GAPP identified similar numbers of the original TPs and FPs, although this varied according to lab. For example, lab 72079 originally reported 47 TPs and 5 FPs, but using pREV in separate (the now recommended decoy database search strategy) with GAPP, it achieved 41 TPs and 3 FPs, and for lab 14997, originally 36 TPs and zero FPs were reported, but GAPP achieved 32 TPs and 1FP. This is an encouraging outcome, as the ABRF data submitters were able to leverage their expertise through manual interpretation, whereas the results in this chapter were generated automatically.

4.5.3 Differences in FPR between data submitters was caused by individual sample handling

Given that the combination of a decoy database and APS scoring is often purported to eliminate FPs, it may seem surprising that FPs were not totally eliminated; the grand mean average number of FPs over all analyses was 1.8 FPs. A list of all FPs observed across the ten labs for (one instance of) all decoys was collated and analysed by the author; none corresponded to the cRAP (“cee-RAP”) – the Common

⁹⁸ Andrews *et al.*, 2006, ABRF-sPRG2006 study: a proteomics standard, see <http://www.abrf.org/ResearchGroups/ProteomicsStandardsResearchGroup/EPosters/ABRFsPRGStudy2006poster.pdf>

Repository of Adventitious Proteins⁹⁹ - list of contaminants, however, four possible contaminant proteins were identified (shown in red in Table 22). These contaminants were not the most ubiquitous FPs.

Table 22 False positive (FP) protein descriptions taken from results of a single instance of each decoy design. There are 49 FPs in total. Rows in bold are FPs with no functional descriptions. In red are the FPs believed to be common contaminants introduced during sample handling.

⁹⁹ This is a list of sequences in fasta format for the most common contaminant proteins in proteomics. It contains proteins from multiple species and includes common laboratory, dust and contact-related proteins, as well as molecular weight or mass spectrometry quantitation standard proteins.

Ensembl identifier	Description
ENSG0000049618	AT-rich interactive domain-containing protein 1B (ARID domain- containing protein 1B) (Osa homolog 2) (hOsa2) (p250R) (BRG1-binding protein hELD/Osa1) (BRG1-associated factor 250b) (BAF250B).
ENSG0000065883	Cell division cycle 2-like protein kinase 5 (EC 2.7.11.22) (CDC2- related protein kinase 5) (Cholinesterase-related cell division controller).
ENSG0000073712	Pleckstrin homology domain-containing family C member 1 (Kindlin-2) (Mitogen-inducible gene 2 protein) (Mig-2).
ENSG0000075914	Exosome complex exonuclease RRP42 (EC 3.1.13.-) (Ribosomal RNA- processing protein 42) (Exosome component 7) (p8).
ENSG0000078487	Zinc finger CW-type PWWP domain protein 1.
ENSG0000096654	Zinc finger protein 184
ENSG0000104133	Spatacsin (Spastic paraplegia 11 protein) (Colorectal carcinoma- associated protein).
ENSG0000108557	Retinoic acid-induced protein 1.
ENSG0000116254	Chromodomain helicase-DNA-binding protein 5 (EC 3.6.1.-) (ATP- dependent helicase CHD5) (CHD-5).
ENSG0000119812	Protein FAM98A.
ENSG0000121495	Retrotransposed gene: no description available
ENSG0000122034	Transcription factor IIIA (Factor A) (TFIIIA).
ENSG0000128731	Probable E3 ubiquitin-protein ligase HERC2 (EC 6.3.2.-) (HECT domain and RCC1-like domain-containing protein 2).
ENSG0000128881	Tau-tubulin kinase 2 (EC 2.7.11.1).
ENSG0000136327	Homeobox protein Nkx-2.8 (Homeobox protein NK-2 homolog H).
ENSG0000138379	Growth/differentiation factor 8 precursor (GDF-8) (Myostatin).
ENSG0000141837	Voltage-dependent P/Q-type calcium channel subunit alpha-1A (Voltage- gated calcium channel subunit alpha Cav2.1) (Calcium channel, L type, alpha-1 polypeptide isoform 4) (Brain calcium channel I) (BI).
ENSG0000141968	Proto-oncogene vav.
ENSG0000143520	Filaggrin family member 2
ENSG0000143702	Centrosomal protein of 170 kDa (KARP-1-binding protein) (KARP1-binding protein).
ENSG0000146872	Serine/threonine-protein kinase tousled-like 2 (EC 2.7.11.1) (Tousled- like kinase 2) (PKU-alpha).
ENSG0000148842	Metal transporter CNNM2 (Cyclin-M2) (Ancient conserved domain- containing protein 2)
ENSG0000152670	Probable ATP-dependent RNA helicase DDX4 (EC 3.6.1.-) (DEAD box protein 4) (VASA homolog).
ENSG0000153201	E3 SUMO-protein ligase RanBP2 (Ran-binding protein 2) (Nuclear pore complex protein Nup358) (Nucleoporin Nup358) (358 kDa nucleoporin) (p270).
ENSG0000156222	Sodium/nucleoside cotransporter 1 (Na ⁺)/nucleoside cotransporter 1) (Sodium-coupled nucleoside transporter 1) (Concentrative nucleoside transporter 1) (CNT 1) (hCNT1).
ENSG0000161849	Keratin type II cuticular Hb4 (Type II hair keratin Hb4) (Keratin-84) (K84).
ENSG0000162896	Polymeric-immunoglobulin receptor precursor (Poly-Ig receptor) (PIGR) (Hepatocellular carcinoma-associated protein TB6) [Contains: Secretory component].
ENSG0000163214	Putative ATP-dependent RNA helicase DHX57 (EC 3.6.1.-) (DEAH box protein 57).

ENSG00000164574	Polypeptide N-acetylgalactosaminyltransferase 10 (EC 2.4.1.41) (Protein-UDP acetylgalactosaminyltransferase 10) (UDP- GalNAc:polypeptide N-acetylgalactosaminyltransferase 10) (Polypeptide GalNAc transferase 10) (GalNAc-T10) (pp-GaNTase 10).
ENSG00000166508	DNA replication licensing factor MCM7 (CDC47 homolog) (P1.1-MCM3).
ENSG00000168924	Leucine zipper-EF-hand-containing transmembrane protein 1, mitochondrial precursor.
ENSG00000169509	Protein NICE-1.
ENSG00000170748	Testis-specific heterogeneous nuclear ribonucleoprotein G-T (hnRNP G- T).
ENSG00000171433	Glyoxalase domain containing 5
ENSG00000171444	Colorectal mutant cancer protein (Protein MCC).
ENSG00000175756	Aurora kinase A-interacting protein (AURKA-interacting protein).
ENSG00000175920	Protein Dok-7 (Downstream of tyrosine kinase 7).
ENSG00000176825	No longer in the Ensembl database
ENSG00000177843	No longer in the Ensembl database
ENSG00000179981	Teashirt homolog 1 (Serologically defined colon cancer antigen 3) (Antigen NY-CO-33).
ENSG00000180043	Pseudogene: no description available
ENSG00000186543	CDNA FLJ41343 fis, clone BRAWH2001973.
ENSG00000188013	Meis1, myeloid ecotropic viral integration site 1 homolog 3 isoform 2
ENSG00000188153	Collagen alpha-5(IV) chain precursor.
ENSG00000188483	Immediate early response 5-like
ENSG00000189182	Keratin, type II cytoskeletal 1b (Keratin-77).
ENSG00000197582	Glutathione peroxidase 1 (EC 1.11.1.9) (GSHPx-1) (GPx-1) (Cellular glutathione peroxidase).
ENSG00000197594	Ectonucleotide pyrophosphatase/phosphodiesterase family member 1 (E-NPP 1) (Phosphodiesterase I/nucleotide pyrophosphatase 1) (Plasma-cell membrane glycoprotein PC-1) [Includes: Alkaline phosphodiesterase I (EC 3.1.4.1); Nucleotide pyrophosphatase
ENSG00000198854	Skin-specific protein 32.

In virtually all cases, a FP was seen by multiple decoys, but by only one lab, implying that the FP was derived from individual sample handling or instrumental/protocol error on a one-off basis (Figure 41 and Figure 42).

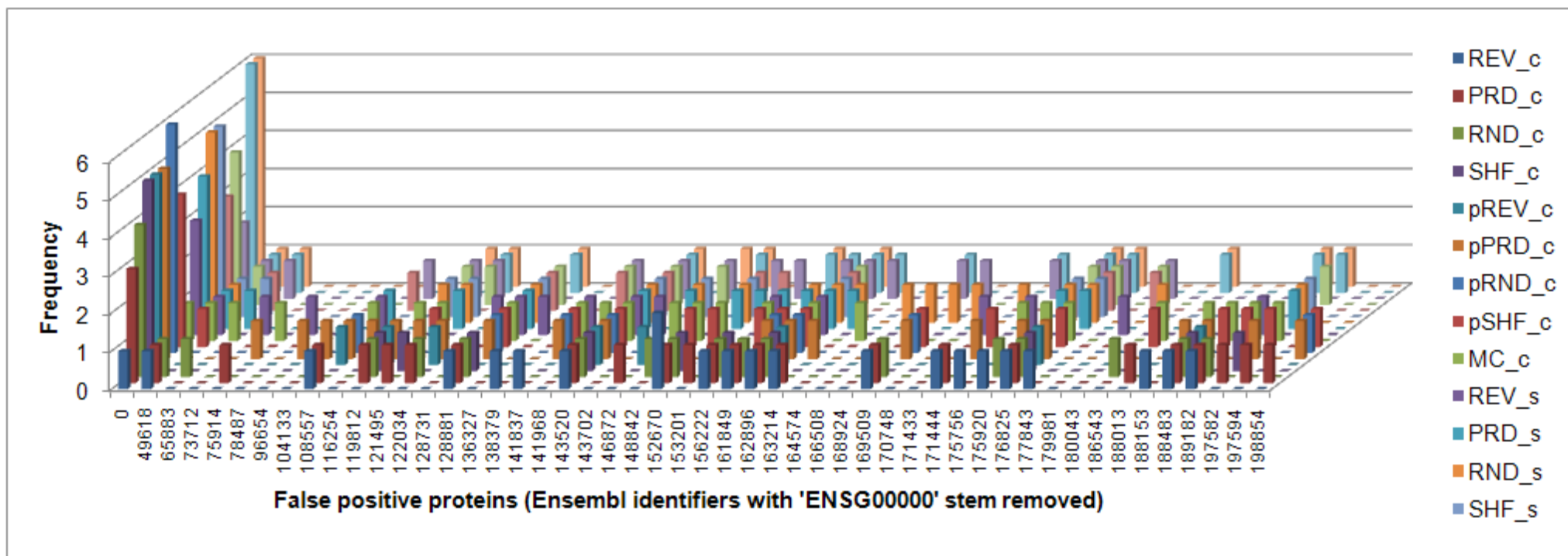


Figure 41 The distribution of FP protein identifications across specific Ensembl gene accession numbers, showing the breakdown by decoy type. The theoretical maximum is 10 for each coordinate on the graph, because each position shows the number of different data submitters, where the FP was found for decoy. The data shown is derived from examination of one instance (out of the ten performed).

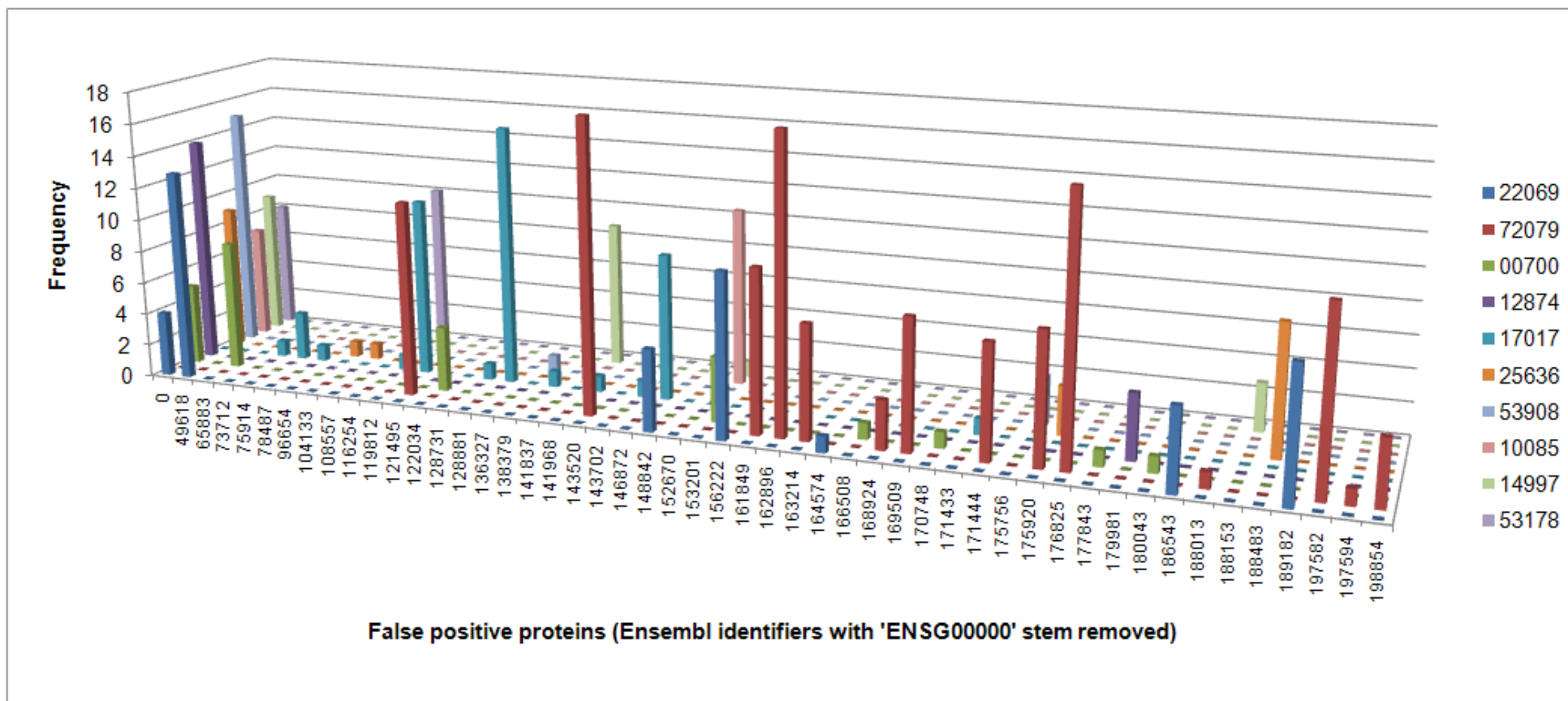


Figure 42 The distribution of false positive protein identifications across specific Ensembl gene accession numbers showing the breakdown by data submitter. The theoretical maximum is 18 for each coordinate on the graph, because each position represents FP observe across 18 different decoy types for decoy instance 1. The data shown is derived from examination of one instance (out of the ten performed).

Tau-tubulin kinase 2 (ENSG00000128881), for example, was observed by 16 decoy designs for lab 17017 (LTQ-FT). Furthermore, lab 72079 (LCQ) had the most FPs by a large margin, suggesting generally poor sample handling; this observation was consistent with the original ABRF study (Andrews *et al.*, 2006).

Significant differences in FPR were detected between instrument types, except for LTQ and HCT, which did not report a significant difference (both being ion traps) (Table 23 and Figure 43).

Table 23 Three-way ANOVA analysis. Given the number of replicates available (a), the least significant differences were found (b), and applied to determine whether significant differences had been observed (c). It showed that there were statistically significant differences between the mean protein FPR values achieved by the instrument types tested (underlined and bold). The only exception was LTQ/ HCT, which did not report a significant difference in FPR. This is to be expected, since LTQ and HCT are both ion trap instruments, and both have the same mass tolerances applied for the protein identification search.

(a)

	HCT	LCQ	LTQ	LTQ-FT	QTOF
No. of replicates	18	18	18	72	54

(b)

	18 to 18	18 to 54	18 to 72	54 to 72	72 to 72
LSD at 5%	2.131	between 2.131 and 1.685	1.685	between 1.066 and 1.685	1.066

(c)

Instrument type		HCT	LCQ	LTQ	LTQ-FT	QTOF
	Mean FPR (prot)	6.04	11.7	6.05	3.84	2.16
HCT	6.04	x	<u>5.66</u>	0.01	<u>2.2</u>	<u>3.88</u>
LCQ	11.7		x	<u>5.65</u>	<u>7.86</u>	<u>9.54</u>
LTQ	6.05			x	<u>2.21</u>	<u>3.89</u>
LTQ-FT	3.84				x	<u>1.68</u>
QTOF	2.16					x

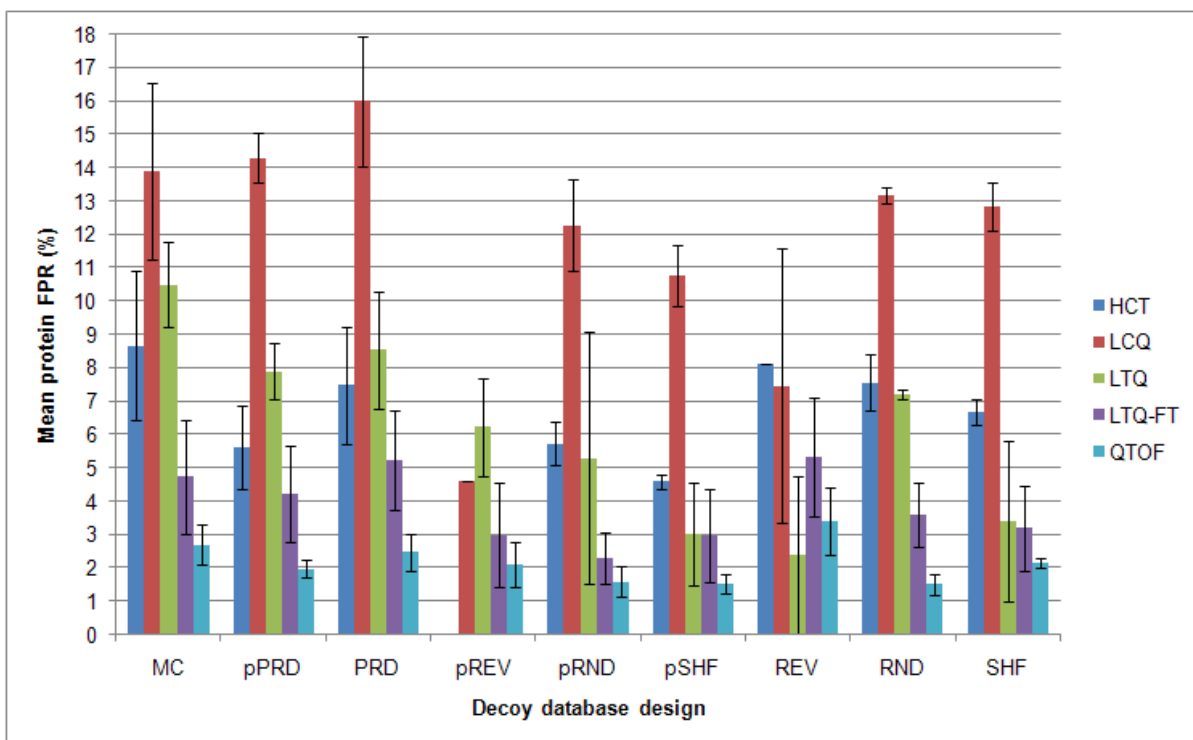


Figure 43 False positive rate by decoy design showing the effect of instrument type. Separate and composite runs are considered together. HCT, LCQ and LTQ were represented by only one replicate for each run type, whereas LTQ-FT was represented by four labs and QTOF by three. The highest resolution instrument, the LTQ-FT did not result in the lowest FPR; although the lack of multiple submissions from a QTOF make its superiority inconclusive.

It should also be noted that only one replicate was present for HCT, LCQ and LTQ. The box whisker plots illustrate the large variation in FPR (Figure 37), which is to be expected since not only the instrument differs, but also the experience and skill of the practitioner.

4.6 Conclusions and recommendations to increase confidence in automated searches

Given the work presented, the peptide level reverse decoy searched independently from the target may be recommended as a suitable alternative to reduce FPs in automated searches. The research offers science-based evidence that may prove useful for guiding proteomic data reporting policies. For example, the recently

released “Paris Guidelines” specify that FPR should be estimated for large scale MS studies using randomised decoy database searches when reporting the findings in publications (Bradshaw *et al.*, 2006, Tabb, 2008). The work here suggests which randomised decoy should be applied. This research also provides a specific example of how publicly available MS/MS datasets may be exploited for novel studies by the community. However, it should be noted that although experimental variance has been covered by use of the ABRF datasets, a specific pipeline was employed for the analysis. As such, the author cannot claim that the observed pattern of decoy performance will definitely be seen in all proteomics workflows. Thus, an investigation with other pipelines, using the presented methodology, would represent valuable work. In particular, it is most likely that the findings will be reproducible across pipelines that employ search engines set to accept *only one peptide identification per spectrum*, as was the case for X!Tandem. X!Tandem did not allow second, third or fourth (etc.) ranked peptides to be found as ‘hits’; only one peptide was successfully identified per spectrum – the top ranking peptide. This is a critical point, because without competition for this top hit position, differences in performance between decoys would most probably have been negligible. In this study, however, separate searches produced lower FPRs than with the composite because there was competitive pressure for the highest score. Mascot, for example, can be set to accept only the top scoring peptide per spectrum in this way.

A research article has been written by the author describing the investigation performed in this chapter. It has been published in the Journal of Proteome Research (see Appendix IV).

4.7 Additional future work

4.7.1 Testing further standard datasets

To further confirm the conclusions made in this chapter, it is proposed that an additional standard MS/MS dataset be applied to the GAPP/ decoy database analysis. There are several candidate datasets that could be applied; Appendix IV lists the alternatives. This work would exploit the existing framework for analysis, with the decoy databases already being generated and Matlab scripts requiring minimal recoding.

4.7.2 Apply peptide level reverse decoy database in the public GAPP pipeline

Now that the peptide level reverse decoy database is recommended for GAPP, this decoy should become the default database for the public GAPP system. At the moment, the protein level reverse is still in use. Ideally, there should be an option for users to select their decoy database of choice, at the time of data submission. The scripts for generating the decoy database instances could be applied to the website to achieve this.

4.7.3 Investigation into theoretical FPRs would make the study more useful for non-standard datasets

FPR was the primary measure for protein identification performance. This is a useful metric, but there are other approaches that could improve the thoroughness of the work. For example, one of the reviewers of the decoy design manuscript suggested that theoretical FPR should be varied to make the results more applicable

to non-standard datasets. The proposed work would involve calculating theoretical FPR - using “classical” approaches (Elias and Gygi, 2007) (Jones *et al.*, 2008a) using $FPR = 2 * \text{decoy hits} / (\text{decoy} + \text{target hits})$, or similar variants, and comparing these values to the real FPRs already obtained. This would mean dropping the APS threshold value to take in more “hits”, and calculating theoretical FPRs at 1%, 5%, 10% for example (in the study already performed all results were based on a single nominal 0% threshold). This additional experiment would be possible because the APS data was stored for all GAPP analysis runs performed in the original study, so the APS threshold can be varied incrementally when querying out the results to obtain plus or minus the actual APS score used as the theoretical 0% FPR threshold.

This work would be informative, because:

- It would highlight the differences a user would obtain from searches where they do not know the identifications *a priori*; as is normally the case. For example, if a search is run versus Ensembl database at a nominal FPR of 5% (measured from decoy hits being FPs) then what would be the real FPR?
- It may highlight further the differences observed between decoy designs, and could confirm the initial findings of the original work.
- It would evaluate performance at a variety of thresholds, as is the norm for ROC analysis. Importantly, there are potential dangers in drawing inference from a single threshold, as was the case in the original study.

In addition, it is proposed that posterior error probabilities (PEPs) could be estimated (Käll *et al.*, 2008), as this approach can be thought of as a measure of local FDR for specific proteins rather than the dataset as a whole. This metric could be calculated as part of the re-analysis of the data presented, and could confirm the findings that the reverse peptide decoy is the recommended choice.

MRMaid, the web-based tool for designing multiple reaction monitoring (MRM) transitions

“For a scientist, all of this is clear, but for a computer it isn't”

Henning Hermjakob, EMBL annual report 2007-08

5.1 Summary

Multiple reaction monitoring (MRM) exploits MS/MS to quantify selected proteins of interest, such as those previously identified in differential studies. Using this technique, the specificity of precursor-to-product ion transitions is harnessed for quantitative analysis of multiple proteins in a single sample.

The design of transitions is critical for the success of MRM experiments, but predicting signal intensity of peptides and fragmentation patterns *ab initio* is challenging given existing methods. This chapter delivers a new tool, MRMAid (pronounced “mermaid”), which offers a novel alternative to streamline the design of MRM transitions, and is aimed at the lab-based proteomics researcher. By exploiting available knowledge of the MRM technique, and using publicly available software programming resources (such as LAMP and PICR), this new tool offers fast and reliable transition design, negating the need for theoretical prediction of fragmentation and removing the need to undertake prior ‘discovery’ MS studies.

5.2 Basic characteristics of an SRM assay

For SRM, a peptide precursor ion that is pre-selected undergoes a CID reaction in MS/MS, generating fragments referred to as product ions. If the chosen pair of precursor and product ion m/z signals is distinct from other m/z signals in the MS run, then it is suitable for monitoring. If this precursor represents a proteotypic peptide (PTP) – one that is unique to the protein of origin and usually visible in MS/MS – then this transition is not only distinguishable from others, but it is a characteristic signature for the protein of interest. Thus, by introducing a heavy surrogate of the native peptide into the reaction, the quantity of the protein target may be accurately determined. Moreover, many peptides may be monitored this way, one after another; hence MRM is a method suitable for multiplexing where several targets may be quantified in the same experiment.

5.3 There is no best practice for designing SRM transitions

The critical part of performing SRM is designing suitable transitions to monitor the protein(s) of interest. In early studies, there was no alternative but to perform empirical ‘discovery’ MS/MS studies to design SRM transitions (Table 24). And even by the early 2000s, no widely accepted best practice had emerged for transition design, and the level of detail given for the selection of candidates was variable in the literature.

Table 24 Overview of empirical methods for SRM transition design applied in the literature to date.

Paper	Year	Study	Transition design criteria and/or method	Software used for S/MRM transition design
(Barr <i>et al.</i> , 1996)	1996	SRM	Peptides chosen because they responded well in FAB-MS.	No
(Barnidge <i>et al.</i> , 2003)	2003	SRM	Peptide chosen by optimising for signal intensity in MS .	No
(Gerber <i>et al.</i> , 2003)	2003	MRM including PTM detection	Peptides chosen based on amino acid sequence and the protease used.	No
(Zhang <i>et al.</i> , 2004)	2004	Distinguished protein isoforms using MRM	Chose two signature peptides based on the four most intense MS/MS TIC peaks for each isoform.	No
(Kuhn <i>et al.</i> , 2004)	2004	MRM	No information given.	No
(Beynon <i>et al.</i> , 2005)	2005	QconCat	Peptides chosen if: did not contain C; were unique in sequence with respect to the other peptides monitored; mass 1000-2000 Da (for MALDI) ; gave strong signal in MS; and had R at terminus.	No, but software was used to predict the gene sequence for the construct.
(Unwin <i>et al.</i> , 2005)	2005	MRM for PTM detection (MIDAS)	Manual calculation and software script developed by Applied Biosystems. MRM Builder software generated precursor-to-product ion transitions and CEs for phosphopeptides. MS used the results to perform sequential scan cycles and select suitable peptides.	Software script developed by Applied Biosystems MRM Builder software for pre-MIDAS workflow
(Cox <i>et al.</i> , 2005)	2005	MRM including PTM detection	Used prototype software with protein sequence and LC conditions as inputs. Performed in silico digest and calculated the m/z of the precursor and fragment ions. Definition of ‘appropriate’ not explicit, but it had to contain S/T or Y. It calculated Q1 and Q3 m/z values for various charge states and fragment ions. The software created multiple possible transitions for the given protein of interest.	‘Research-grade’ version of software from the Applied Research group at MDS Sciex
(Ciccimaro <i>et al.</i> , 2006)	2006	MRM for PTM detection	Most intense precursor ion and product ion chosen. Applied MIDAS approach for prediction and data acquisition.	MIDAS software
(Anderson and Hunter, 2006)	2006	MRM	Predicted tryptic peptides , found corresponding Swissprot annotation, found physico-chemical parameters of each (composition, mass, Hopp-Woods hydrophobicity annotation (Hopp and Woods, 1981)), predicted RT (Krokhin <i>et al.</i> , 2004). Determined likelihood of detection of each peptide using a published plasma data and calculated each protein: (no. of hits for peptide)/(no. of hits for most freq detected peptide for that protein) and index of protein quality by positively weighting: P, KP, RP, DP and negatively weighting: C, W, M, chymotrypsin sites, detrimental Swissprot features and mass less than 800 and more than 2000. Length set at 8-24 amino acids. Used GPMDB to selected peptides from frequently detected proteins.	Used in silico techniques including software tools to predict physico-chemical properties. Used GPMDB repository to confirm proteotypic peptides from MS results (Craig <i>et al.</i> , 2004). No integrated software program used.
(Wolf-Yadlin <i>et al.</i> , 2007)	2007	MRM for PTM, used iTRAQ labeling not radio-isotope labels	Performed discovery study. Considered peptide m/z and charge state, the characteristic b- and y-ions and CE required.	No
(Stahl-Zeng <i>et al.</i> , 2007)	2007	MRM using elution time constraints to maximise number of transitions	Used criteria of the proteotypic species (as defined by PeptideAtlas) including that the peptide ionises well in the range of the MS instrument and is unambiguously associated to a single protein.	Used bespoke software modifications for instrument control and data acquisition in the MRM study.

Paper	Year	Study	Transition design criteria and/or method	Software used for S/MRM transition design
(Kay et al., 2007)	2007	MRM	Some transitions taken from papers, others chosen using criteria: most intense y-ions, two product ions were monitored, extracted chromatograms for peptides of interest and used HPLC peaks at identical RT for the two transitions. When HPLC peaks were summed signal-to-noise ratio was required to be a minimum of five.	No
(Keshishian et al., 2007)	2007	MRM	Factors considered were: observed or predicted RT, MW, and charge state. Preference was given to moderately hydrophobic peptides likely to produce triply or doubly charged ions in detectable mass range.	No
(Lenz et al., 2007)	2007	MRM	Decision was based on collision energy (in Q2), specific m/z and dwell time required.	No
(Lange et al., 2008)	2008	MRM	Predicted transitions using TIQAM.	TIQAM

5.3.1 Summary of existing transition design software

Despite the lack of consensus, various commercial, vendor-specific software packages for MRM transition design are available, including, for example MIDAS™ (MRM-initiated detection and sequencing) Workflow Designer software (ABI, Foster City, CA), which calculates theoretical peptides and corresponding transitions, then builds the MIDAS acquisition method (Unwin *et al.*, 2005), whereby a coupled Q-TRAP (ABI) iteratively cycles through scans and to select suitable peptides. Thermo Scientific's (Waltham, MA) contribution has been presented to user groups but was not available for purchase at the time of writing. It takes data from SIEVE, the Automated Label-Free Differential Expression Software, computes ions generated, relates those back to the sequence and then imposes filters similar to those in MRMAid (see later section).

In 2007, however, when the author was developing MRMAid, there were no publicly available software programs to predict suitable transitions. In 2008, at the same time as MRMAid was released, TIQAM (Targeted Identification for Quantitative Analysis by MRM) (Lange *et al.*, 2008) was published by the Institute of Systems Biology, Seattle. This was the first tool to mine a public MS repository, PeptideAtlas (Deutsch *et al.*, 2008), for peptides based on the number of previous observations. It fundamentally differs from the approach taken by MRMAid, because it requires the user to experimentally acquire MS/MS data to isolate suitable candidates from the list of possible transitions, whereas MRMAid is able to indicate which peptides in the shortlist are most suitable using its novel transition scoring algorithm. MRMAid is

also a web-based tool whereas TIQAM is designed to be installed locally, requiring setup of a local database.

In summary, at the time of development there were limited options available for MRM practitioners requiring transition design support without involving acquisition of MS/MS data for the prediction process. In this chapter, a publicly accessible MS vendor-independent program, called MRMAid, is presented, which provides transition design support to assist the expansion in use of the MRM method.

5.4 Methods

5.4.1 An overview of the MRMAid system

MRMAid's method for transition design relies on the combination of two sources of information: firstly, on prior knowledge of the kind of precursor peptides that generally perform better in MS/MS, and secondly, on mining the data in a proteomic data repository, GAPP (Shadforth *et al.*, 2006), to determine which precursor and corresponding fragment ions have appeared regularly for the given protein of interest. The hydrophobicity and reverse phase RT of each peptide candidate are calculated so suitable transitions may be selected and ordered. Finally, as with all approaches for transition design, the shortlist of the best transition candidates must be validated using a suitable MS instrument.

The GAPP database is populated with data from public resources and data submitted directly by users, as described previously in this thesis. Peptide candidates for MRM are mined from this data source by applying the principle of proteotypicality (Mallick *et al.*, 2007), whereby peptides that map unambiguously to a single protein are first mined from the database.

To design an MRM experiment, several SRM transitions are monitored in a single assay. For this purpose, MRMAid allows comparison of individual SRM transition candidates using estimated RT and hydrophobicity values, this way transitions may be selected to avoid co-elution of peptides. The proposed peptide candidates can be compared by downloading the page of MRMAid results and comparing RT and transition score (TS) values. This process is explained below, and ultimately allows users to design the optimal bespoke MRM experiment for their specific target proteins.

To indicate reliability of the transitions, metrics of reproducibility are calculated for the candidate product ions. An indication of reproducibility is required because GAPP is a public system, and as such, accepts data from any source if the format and metadata requirements are met. Reliability is ensured in two discrete ways: firstly, peptide precursor (and subsequent fragment) candidates are presented to the user with the number of times they have been observed in GAPP for the protein of interest. Secondly, the individual product ions are assessed in terms of signal intensity reproducibility: an average of signal intensity for the relevant m/z peaks is

calculated across all applicable experiments, as well as the variance and standard deviation values. These descriptive statistics indicate the reproducibility of the fragment ions for a given precursor, and hence, point to the number of times one would expect to have to run the MRM experiment to observe a good result for the transition.

5.4.2 Software implementation

MRMaid (Figure 44) is a program written in Perl and PHP that interrogates the GAPP database for suitable transitions for a given protein sequence. The protein target is input by the user as a database accession number, such as an Ensembl, Swissprot or IPI number. Many database accession numbers are supported thanks to integration of the EBI's PICR service (Côté *et al.*, 2007).

The steps in the algorithm for transition design include the following:

- (a) All PTPs for the protein are retrieved
- (b) PTPs are filtered by criteria defined by the user
- (c) MS/MS data for the protein is retrieved, descriptive statistics are calculated and γ - and b -ions are assigned for each peptide using the mass tolerance window provided by the original data submitter
- (d) TS value is computed for each peptide and is used to rank the transitions
- (e) RT and hydrophobicity are computed for each peptide

The following sections explain the MRMaid workflow (Figure 44) in more detail. The algorithm is novel and was conceptualised and formulated by the author. With help and direction from the author, Vanessa Ottone (visiting student) implemented the function to assign b - and γ -ions using mass data.

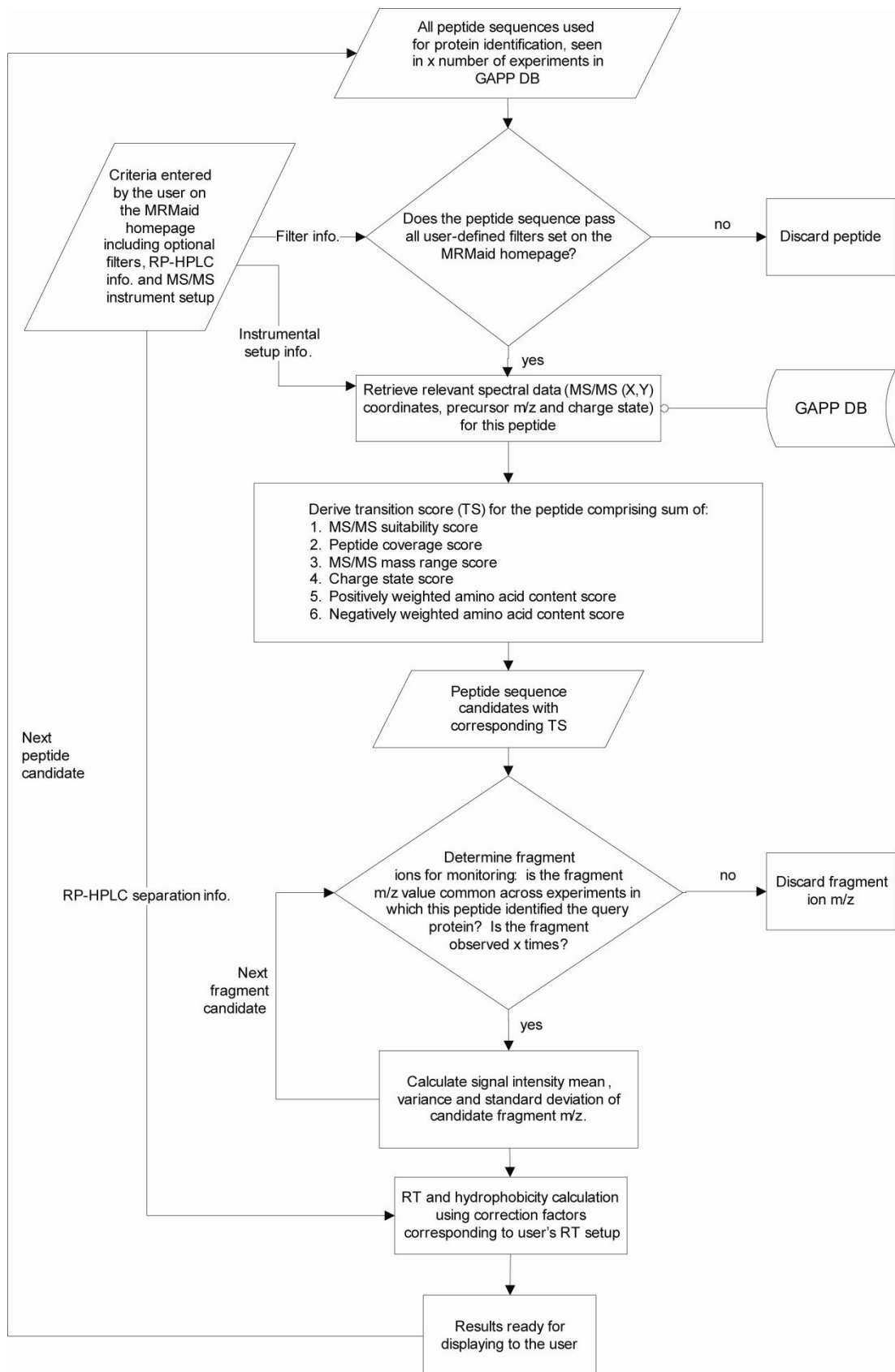


Figure 44 The process of transition design in MRMAid. Transition score is used to rank the resulting transition predictions. x is a value that can be adjusted according to users' requirements, in the web release it is one.

Luca Bianco implemented the PHP script for the MRMAid website homepage, using Darren Oakley's original GAPP style-sheet. He also implemented the PHP program (using the GD library) to draw the spectral graphics.

5.4.3 MRMAid has filters to determine optimal transition candidates

Users may choose from a series of filtering options to constrain the prediction of transitions (Table 25). These filters take the form of a list of check boxes, drop down menus and text boxes on the homepage. They can be found below the box where the user enters the accession number for the protein target. The filters were chosen through targeted questioning of experts in the MRM approach and the use of prototypes for live demonstration. The author organised meetings and lead the sessions with expert practitioners from Quotient BioResearch Ltd., a contract research organisation, and the University of Cambridge. The reasoning given for inclusion of the criteria suggested by the experts is described below and is summarised in Table 25. All peptides that pass the relevant filtering steps enter the transition scoring phase, which is necessary to rank the candidates.

Table 25 Optional filters that may be used to constrain the search for MRM transitions using MRMAid

Filter criterion	Description
Peptide observation redundancy	Peptide must have been seen in x% of all observations of the query protein in GAPP database
Internal cleavage	Peptides with K or R, unless followed by P
Instrument type	Constrains to MS/MS data retrieved for the instrument set up selected
Omit N and Q	N and Q can be deamidated resulting in m/z reproducibility issues
Peptide length	Short peptides (<7 or 8 amino acids) are unlikely to be unique to the target, and very long peptides will be out of mass range
Omit M and C	Often covalently modified affecting m/z
Omit Q and E	Can spontaneously cyclise to form pyroglutamate
Accept only P-containing peptides	Produces a very high abundance peak – suitable when a single product ion is sufficient, such as in low complexity samples
Omit P at any position	Swamps tandem spectrum, selected when multiple transitions per target are required, such as in high complexity samples like serum
Omit P ₁ (P adjacent to the C-terminus)	Can produce non-specific product y-ion
Omit P ₂ (P second position from C-terminus)	Can produce non-specific product y-ion
m/z cut off (user specifies a value, x)	Selects only fragment ion m/z > x, where x is a percentage of the precursor m/z

Users may define the proportion of times a peptide has been observed for the target protein in GAPP database. In this way, the frequency of peptide identifications in the repository may be used as a measure of transition reliability for the protein. For example, if users enter '50' then this means that the peptide candidate(s) presented in green in the peptide results table (see Figure 45) are assigned in at least 50% of occasions when the protein target was successfully identified in GAPP, so these are the best candidates to choose. The higher the value entered, the more stringent the prediction.

As a default, peptides with internal cleavage sites (namely peptides with K or R, unless followed by P) are omitted to prevent selection of peptides that may be

irregularly cleaved. This is a necessary feature since the efficiency of trypsin is known to be only (approximately) 70 percent (Yen *et al.*, 2006, Falth *et al.*, 2007).

Another important consideration for users intending to perform MRM is whether MS/MS evidence is available for their particular MS instrument. Each instrument is known to have a different set of preferred PTPs (Mallick *et al.*, 2007), therefore, to account for this phenomenon, GAPP database's MS instrument information may be incorporated as a filter into the querying process. A drop-down menu to select the type of instrument is provided on the homepage.

Protein expression can vary significantly between tissue types. To try to account for this, the type of biological sample can also be specified as a filter in the search for transitions. In serum, this is particularly useful, since its levels of protein expression and sample complexity can present unique challenges for the transition design. The ability to choose transitions based upon experiments performed on the same type of sample increases the likelihood of selection of successful candidates.

In addition to constraining experiment-specific factors, information on the sequence of the peptide may also be used in the filtering process. For example, N and Q may be deselected, because these residues can be deamidated resulting in fragment ion m/z irregularities and problems with reproducibility. This is a problem because the m/z of fragment ions must be as consistent, and hence reliable, as possible. Opting to

omit peptides containing Q or E at the N-terminus is also possible in MRMAid, because these residues can spontaneously cyclise to form pyroglutamate. Likewise, P is an important residue to consider when designing transitions. Peptides containing P may be considered favourable, because they generally produce MS/MS peaks of high intensity. This is because proline's 3D structure promotes fragmentation, often producing a single greater abundance fragment ion that 'swamps' the remainder of the tandem mass spectrum. However, this single signal may not be sufficient to identify the protein with adequate specificity, particularly in complex samples. For this reason, users may opt to omit or allow candidate peptides containing P. The location of P in the peptide primary sequence is also important. P, which is adjacent to the C-terminus (P1) or in the second position from the C terminus (P2), is generally not desirable in MRM. This is because a very short, non-specific product y-ion is produced, which is unsuitable for monitoring. Users may, therefore, opt to omit P1 and P2-containing peptides from their results.

Peptide sequence can also affect the probability of covalent modification. M and C are often modified residues, so if they are omitted it makes it possible to constrain the candidates to those where mass should not vary. Naturally, if all possible amino acids that fall into this category were omitted, the filter would be too strict, however by negatively weighting these residues in the TS calculation it provides a more suitable method (see the next section for more detail). Typical modifications like carbomido-methylation are, however, worth considering in peptides for monitoring, as there are examples in the literature, such as for Apolipoprotein A2 (Kay *et al.*,

2007) where they have been present in transitions, so for this users may choose a filter in this list to omit them.

A further filter option is peptide length. Users may constrain this because very short peptides (<7-8 residues) are unlikely to be unique, and peptides longer than approximately 20 to 25 residues are unsuitable for MRM because they may exceed acceptable mass range. For all MRMAid searches, mass range for the peptide (MS mode) is restricted to 500-1600 *m/z*. This range is routinely used for monitoring tryptic peptides, so peptide candidates beyond this range are omitted by default.

b- and y-ions are common fragment ions produced in tandem MS. To restrict transitions to peptides that produce suitable y- or b-ions for monitoring, two approaches are employed in MRMAid. Firstly, y-ions (shown in red) and b-ions (in blue), with the charge states, are highlighted in the resulting spectrum graphic and results table (an example is shown in Figure 45).

Peptide candidates for ENSG00000118271 ?

Peptide ?	Observations ?	Avg TS ?	m/z to monitor ?	Hydrophobicity ?	Retention time ?
GSPAINVAVHVF R	72	39.41	547.029 - 1789.89	33.6	23.8 min
AADDWEPFASG K	84	38.69	558.001 - 1966.1	19.39	18.3 min
ALGISPFHEHAEVVFTANDSG PR	29	25.88	982.137 - 1999.82	38.79	25.8 min
YTIAALLSPYSYSTTAVV TNPK	40	24.15	945.08 - 1992.07	44.02	27.8 min
TSESGELHGLTTEEFV EGYK	52	23.82	662.756 - 1996.6	41.58	26.8 min

Export in TSV ?

(*) Warning: Default selected therefore retention times may be inaccurate

Product ions for ENSG00000118271 ?

Fragments to monitor ?		View b and y ions only				
y11	1+	lon at m/z 1222.29 - 1222.83	seen 16 times	Intensity mean 8.938	Intensity var 468.996	Intensity std dev 21.656
b1	1+	lon at m/z 1191.9 - 1193.41	seen 61 times	Intensity mean 5.639	Intensity var 64.301	Intensity std dev 8.019
y10	1+	lon at m/z 1125.19 - 1126.42	seen 47 times	Intensity mean 4.553	Intensity var 20.513	Intensity std dev 4.529
y9	1+	lon at m/z 1054.29 - 1055.39	seen 57 times	Intensity mean 8.421	Intensity var 35.427	Intensity std dev 5.952
b2	1+	lon at m/z 1045.12 - 1046.4	seen 61 times	Intensity mean 9.213	Intensity var 11.237	Intensity std dev 3.352
b3	1+	lon at m/z 946.144 - 947.266	seen 63 times	Intensity mean 7.063	Intensity var 9.996	Intensity std dev 3.162
y8	1+	lon at m/z 941.195 - 942.24	seen 70 times	Intensity mean 42.429	Intensity var 321.263	Intensity std dev 17.924
y7	1+	lon at m/z 827.068 - 828.269	seen 62 times	Intensity mean 13.629	Intensity var 29.811	Intensity std dev 5.46
b4	1+	lon at m/z 809.051 - 810.234	seen 65 times	Intensity mean 6.569	Intensity var 8.78	Intensity std dev 2.963
y6	1+	lon at m/z 728.153 - 728.617	seen 71 times	Intensity mean 37.775	Intensity var 155.977	Intensity std dev 12.489
b5	1+	lon at m/z 710.042 - 711.3	seen 50 times	Intensity mean 3.5	Intensity var 8.582	Intensity std dev 2.929
y5	1+	lon at m/z 656.676 - 657.675	seen 69 times	Intensity mean 17.087	Intensity var 37.169	Intensity std dev 6.097
b6	1+	lon at m/z 638.679 - 640.106	seen 63 times	Intensity mean 7.937	Intensity var 52.157	Intensity std dev 7.222
y11	2+	lon at m/z 620.12 - 621.613	seen 9 times	Intensity mean 5	Intensity var 11	Intensity std dev 3.317
b1	2+	lon at m/z 596.182 - 597.318	seen 14 times	Intensity mean 2.286	Intensity var 1.758	Intensity std dev 1.326
y10	2+	lon at m/z 571.566 - 572.988	seen 5 times	Intensity mean 6.4	Intensity var 12.3	Intensity std dev 3.507
y4	1+	lon at m/z 558.056 - 558.507	seen 71 times	Intensity mean 31.521	Intensity var 109.796	Intensity std dev 10.478

Experiment 153

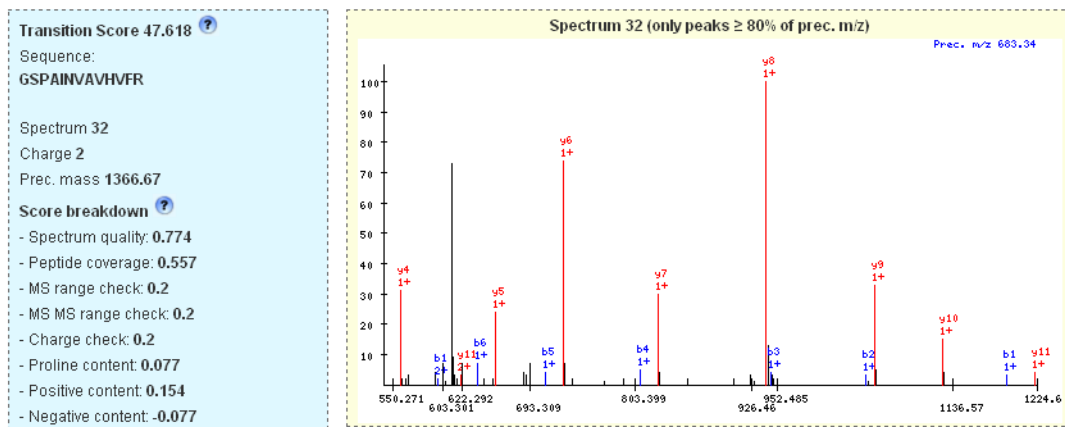


Figure 45 MRMAid has three main views for the candidate transitions. This shows screenshots of elements of the interface for transitions predicted for transthyretin (ENSG00000118271). The first (top) is the list of peptide candidates ordered by transition score (TS). ‘m/z to monitor’ is the range within which the product ion masses fall. The peptide sequences are hyperlinks, when a user clicks on ‘GSPAINVAVHVF’, for example, the product ion results are displayed (centre) and a list of schematic spectra (bottom) that represent the underlying data in GAPP DB that was used for the predictions. The ‘Export in TSV’ link (top) allows users to download all the peptide and product ion data for the target.

y- and b- ion assignment is achieved by dynamic computation using the accepted rules of fragmentation (Zhang *et al.*, 2005), and where the theoretical m/z values (for monoisotopic and average mass - as required) are compared to those observed in the underlying MS/MS data. The mass tolerance window is applied to ensure the resolution of the data is accounted for in the assignment; for this, the range specific to each experiment in GAPP is used, and not the maximum over all relevant observations in the search. In an extremely small number of cases, this window prevents unambiguous assignment of a peak to a single b- or y-ion. In this case, the ion is not counted as b- or y- and is instead grouped into the 'other ion type' category (shown in black).

The second approach to ensure MRMAid suggests suitable fragment peaks is by using m/z cut-off; an approach that is demonstrated in the literature (Anderson and Hunter, 2006, Keshishian *et al.*, 2007). This facilitates selection of fragments with m/z greater than the precursor m/z . If desired, only MS/MS fragment masses that are x percent or more of the mass of the precursor will be recommended by MRMAid, where x is specified by the user. For example, if x is 100%, then only fragment ions having a mass higher than the precursor ion will be selected. This ensures that fragments in the higher end of the m/z spectrum will be considered and accounts for the effect of the mass filtering step that is applied between MS and MS/MS modes. This approach has the knock-on effect of increasing the chances of producing a more reliable and specific fragment candidate, because the specificity of a transition greatly increases when looking at a product ion that is higher m/z than the precursor

ion. Note that there is no limit to the number of fragment ions that can be suggested to the user – only the underlying data restricts this.

5.4.4 MR Maid takes a novel approach to transition ranking

MR Maid's transition score (TS) provides a quantitative measure of predicted performance in MRM to candidate transitions that are retrieved by mining the GAPP database and by aggregating the results. Implementation of a series of Boolean filtering steps (as in the filters described above) was one available options for this part of the process. However, this would not reflect the approach taken by experts, and would have made it impossible to judge relative favourability of candidate transitions. Therefore, TS is calculated as the weighted sum of several key characteristics of the spectral data (each denoted as a letter: q , c , r , s , p and n), giving a quantitative measure of expected performance in MRM (Table 26). The relative weights applied (50, 8, 7, 6, 1 and 1, respectively) were determined by combining authors' experience with researching the literature and discussions with practising MRM experts. Note that all values are normalised to a range between zero and one, such that each coefficient has consistent scale.

Table 26 Derivation of Transition Score (TS) coefficients. TS is used to rank the predicted transitions in MRMAid. It is calculated as the sum of the coefficients, each relating to efficiency in MRM. Each coefficient is derived by multiplying the value by the weighting.

Transition score coefficient	Description	Relative weighting
MS/MS suitability (q)	Assesses MS/MS spectrum for suitability in MRM by assessing m/z values of y -ions	50
Peptide coverage (c)	Favors fragment ions that represent a greater proportion of the precursor peptide sequence	8
Mass range (r)	Constrains acceptable MS/MS mode mass range	7
Precursor charge state (s)	Doubly and triply charge peptides are favored	6
Positively weighted residue content (p)	H, K and R	1
Negatively weighted residue content (n)	Y,S,T,C and M	1

The MS/MS suitability score (q) positively weights precursor peptides that demonstrate a suitable profiles of y -ions for MRM. A suitable y -ion profile is one where there are several y -ions at the high m/z range of the MS/MS spectrum. y -ions with a 1+ charge state were chosen for this because these generally have higher m/z than b -ions, so are a suitable indicator to determine the quality of a spectrum for MRM. They are also the ions that are routinely used, as demonstrated, for example, in Anderson and Hunter's paper (Anderson and Hunter, 2006) where virtually all experimentally confirmed transitions were y -ions: see Table 27, where selected transitions from this article are used to validate MRMAid performance. Furthermore, b -ions are less suitable than y -ions because they are susceptible to cyclisation, which can result in fragment ions of unexpected sequence (Harrison *et al.*, 2006). Therefore, the q coefficient, and hence overall TS value, is designed to favor spectra showing evidence of multiple y -ions and is achieved by considering three key factors for each precursor peptide sequence that passes the initial filtering stage for the protein of

interest. First: the highest m/z value of the MS/MS spectrum that is a y-ion with 1+ charge ($(m/z)_{\max}$); second: the number of peaks of the spectrum that are y-ions with 1+ charge (n_{peaks}); and third: the standard deviation of m/z values of peaks of the spectrum that are y-ions with 1+ charge (std_{peaks}). Standard deviation is calculated in the usual way (Equation 1), where $(m/z)_{peak}$ is the individual m/z value of a y-ion with 1+ charge in the MS/MS spectrum, and $\overline{(m/z)}_{peaks}$ is the mean m/z value of all the individual y-ions with 1+ charge for the given spectrum, for the peptide in question.

Equation 1 Calculating standard deviation, which is used to calculate the coefficient q in MRMAid's transition score

$$std_{peaks} = \sqrt{\frac{\sum ((m/z)_{peaks} - \overline{(m/z)}_{peaks})^2}{n_{peaks} - 1}}$$

The rationale for using the three factors above is the following: if it is possible for a peptide to fragment into n different y-ions, then the MS/MS metric for suitability in MRM should favour spectra with evidence of larger y-ions (those having higher m/z values), as well as a higher number of y-ions in total. Since, in general, the whole complement of theoretically possible y-ions, for a given peptide, are not produced, instead spectra with a smaller standard deviation of m/z values belonging to identified y-ions must be favoured.

The three elements, above, are combined to assign higher scores to spectra carrying evidence of the heaviest y-ion, as well as the greater number of y-ions. If two or more spectra are equal according to these first two criteria, the one identifying a higher number of heavier y-ions gets a higher score. Thus, given a spectrum for a peptide sequence in the GAPP database, its interim score k_r (r denoting 'real') is calculated as shown in Equation 2.

Equation 2 Calculating the interim score k_r as part of coefficient q

$$k_r = \frac{(m/z)_{\max} \cdot n_{\text{peaks}}}{\sqrt{1 + \text{std}_{\text{peaks}}}}$$

After this, using the peptide sequence, a theoretical interim score k_t (t for 'theoretical') is similarly calculated, but this time considering the total complement of theoretical y-ions (with 1+ charge state) for the sequence, using the rules of CID fragmentation (Zhang *et al.*, 2005). To finish, the MS/MS suitability score (q) is computed as a ratio between the real and theoretical interim scores (Equation 3), thus, q is normalised to a scale between zero and one, as is the case of all other coefficients in TS.

Equation 3 MS/MS suitability score (q) is computed as a ratio between the real and theoretical interim scores

$$q = \frac{k_r}{k_t}$$

In summary, q provides a quantitative scale of MS/MS suitability to each peptide based on the MS/MS spectral evidence, and is weighted most heavily because it is quantifying actual experimental data.

Peptide coverage (c) refers to the proportion of the target peptide that is represented by the product ions for the transition. Product ions that have greater m/z are preferable, because they represent a greater proportion of the original peptide in the spectrum and, therefore, increase the specificity of the transition. Peptides with fragments within the mass range for MS/MS mode (r) of 500-1600 m/z are positively weighted, for reasons mentioned earlier.

Most tryptic peptides are doubly charged cations, with one charge originating from the primary amine terminus and one from the K or R residue side chain at the C-terminus. Precursor charge state (s) is important for MRM because doubly or triply charged precursor peptides are favoured, due to the mass filtering stage. Doubly and triply charged precursors, therefore, achieve a coefficient higher than singly charged peptides, such as those derived from unspecific protein cleavage. Charge state for each precursor peptide is known, because this information is available in GAPP's input peak lists (currently .mgf, .pkl or mzXML). Fragment ions with a single charge are preferred in MS/MS mode; information on charge state is displayed as b- and y-ion labels on the peaks, so the user can select singly charged peaks for monitoring.

Weighting of specific residue content (coefficients p and n) allows more refined ranking to be performed on the candidate output list. Positively - charged residues

have been demonstrated to increase fragmentation and hence visibility in MS, so are positively weighted (Mallick *et al.*, 2007). Y, S and T, like C and M, may be post-translationally modified, so are negatively weighted.

5.4.5 Transitions can be scored in the absence of MS/MS evidence

TS may also be calculated in the absence of MS/MS data. In this situation, it is computed with q omitted and the user is informed on the results page. This calculation involves digestion of the protein and application of user-defined filters as in the usual MRMAid mode. Clearly, candidates predicted in the absence of MS/MS evidence are less robust, because predicted peptides cannot be statistically measured for reliability across experimental datasets, and less comprehensive, because candidate MS/MS fragment ions cannot be suggested. However, an indication of which peptides should be monitored is welcome functionality, since the alternative is deciding on transitions manually, which would be tedious for large proteins, and particularly problematic when many proteins are to be analysed.

5.4.6 Results can be downloaded from MRMAid

In the final output, MRMAid provides a tabulated list of ranked transition candidates, which are intended for validation using the MS instrument. This validation is necessary because it is not possible to select a single *in silico*-derived transition, without experimental confirmation, as shown by existing software options, such as MIDAS workflow designer. The results table may be exported as tab-separated values (TSV) for import into a spreadsheet package for local analysis and archiving. Also, as described previously, schematic spectra are displayed to highlight the y- and b-ions suggested for monitoring.

5.4.7 Retention time is calculated using a linear model

Elution time data is commonly used in combination with MS/MS to support transition design for MRM experiments (Stahl-Zeng *et al.*, 2007, Wolf-Yadlin *et al.*, 2007, Keshishian *et al.*, 2007). RT is used in the discovery phase to decide which peptides to monitor - on the basis of peak separation and, once transitions are chosen, it allows transition ordering (Wolf-Yadlin *et al.*, 2007) and provides confirmation that the peptide species expected is the one actually being monitored (Stahl-Zeng *et al.*, 2007, Keshishian *et al.*, 2007, Kay *et al.*, 2007). Management of elution time also maximises the number of transitions that can be performed, without overly compromising on sensitivity (Stahl-Zeng *et al.*, 2007).

Implementation of RT was necessary for MRMAid to become an MRM, rather than an SRM, design tool. By having RT indicators to compare across all candidate transitions, the order of transitions in a multiplexed experiment can be planned. The problem in achieving this is that GAPP is dependent on data sources in the public domain and RT information is rarely captured in publicly available datasets. In the absence of this data, RT is predicted using a linear peptide RT algorithm, which holds true for peptides up to approximately 20 residues (Krokhin *et al.*, 2004). This was acceptable since MRM peptides are typically no more than 24 amino acids. The procedure involves summation of residue coefficients then correction for the length of the peptide and factors relating to the procedure and setup of the column. There are eight different options for reverse phase column setup provided by MRMAid for this process, each of which can be selected by the user using a drop-down menu.

There is a default option available, namely, a microflow method (4 μ l/min) applying a gradient of 1-80% ACN over 60min (1.32% ACN/min) using a 150 μ m \times 150mm Vydac® 218 TP C18 bead column (5 μ m). It is recommended that the user selects the specific setup that reflects his/her own RP chromatography procedure. Therefore, there is a warning printed on the results page when the default option is selected, to remind users that they should choose a specific setup wherever possible. RT prediction is programmed in a modular fashion, so that as improved RT models become available, new algorithms may be integrated into MRMAid. The author implemented the RT prediction program in conjunction with Vanessa Ottone.

5.5 Results

5.5.1 Man versus MRMAid: testing MRMAid's performance

To demonstrate MRMAid's ability to predict MRM transitions, a selection of transitions that had been experimentally validated were obtained by the author and compared to the results generated by the MRMAid program. Table 27 shows the comparison with transitions from Anderson and Hunter (Anderson and Hunter, 2006), and Table 28 transitions provided by Chris Barton of Quotient Bioresearch.

The author obtained diverse datasets from Human Proteinpedia, PeptideAtlas and Tranche and analysed and stored them using GAPP. The data included serum-based identifications required for the test cases. In addition, four MS/MS datasets for horse serum proteins were obtained and submitted to GAPP pipeline (with appropriate search parameters based on the metadata): these were two whole

plasma samples, and two samples depleted of highly abundant proteins using ACN-depletion, taken from eight horses see (Barton *et al.*, 2009) for details. The data, in total, represented 100 runs in QQQ-MS/MS.

Table 27 MRMAid performance versus experimentally verified transitions (Anderson and Hunter, 2006). The results shown are derived from MRMAid searches using the default search parameters, namely, no internal cleavage sites, 80% of the precursor mass and 50% of total observations. Peptides which did not meet the criterion of 50% of total observations criterion (shown as red rows in the table of results) were still considered, due to current quantity of available MS/MS reference data. An ‘observation’ in this case is an identification made using a single MS/MS dataset that was submitted to GAPP. Results are derived from searches performed on 26th August, 2008. Values are rounded to one decimal place in the table but MRMAid shows up to three in the results on the website. Key: (a) these values may be shown as a range to account for the mass tolerance of fragment ions entered by the user when the MS/MS data was submitted to GAPP; (b) relative intensity refers to the fact that signal intensity is normalised by the GAPP analysis pipeline. When spectra are uploaded for protein identification, the y-dimension of each spectrum is normalised to 100. This means that for all the spectra in the GAPP database, which are mined by the MRMAid program, there is a maximum y value of 100. This is also reflected in the graphical MS/MS spectra displayed on the MRMAid product ion results page. The intensity values differ for each individual peak in each experiment submitted to GAPP, therefore the mean over all observations is given for each particular ion species, for example, y8. It is possible to successfully monitor product ions at low abundance, as long as there is no overlap with other peaks: generally there is less likely to be overlap at the very high m/z end of the spectrum, therefore the higher the m/z , the lower the abundance that can be successfully tolerated.

Protein target (Ensembl gene identifier)	No. of peptide candidates found for target	Precursor peptide sequence used by Anderson and Hunter	No. times peptide seen in GAPP for target	Average TS for peptide	No. of predicted (b-ions, 1+ only)	No. of predicted (y-ions 1+ only)	RT real (min)	RT predicted (min)	MS/MS transition real (m/z)	MS/MS transition predicted (m/z) ^a	No. of times product ion seen	Mean relative intensity for product ion observations ^b
Afamin (ENSG00000079557)	9	DADPDTFFAK	9	43.4	5	5	21.5	19.9	825.4 (y7) 940.4 (y8)	825.3 - 825.3 940.2 - 940.4	7 7	100 35.1
Alpha1-acid glycoprot. 1 (ENSG00000187681)	6	NWGLSVYADKPETTK	14	36.6	5	6	19.7	20.7	1052.5 (y9) 1068.5	1052.3 - 1052.5 1068.6	13 1	91.5 7
Alpha-2-antiplasmin (ENSG00000167711)	5	LGNQEPGGQTALK	13	33.7	7	6	12.6	14.9	771.4 (y8)	771.2 - 771.4	13	100
Alpha-1-antitrypsin (ENSG00000197249)	18	DTEEDFHVDQVTTVK	55	29.9	9	8	17.4	21	790.4 (y7) 889.5 (y8)	789.9 - 791.4 888.7 - 890.0	53 53	42.3 96.8
Alpha-2-macroglobin (ENSG00000175899)	47	LLIYAVLPTGDVIGDSAK	22	29.1	11	8	36.5	28.7	1059.5(y11) 1172.6(y12)	1059.1 - 1059.6 1172.2 - 1172.7	21 20	99.3 25.9
Angiotensinogen (ENSG00000135744)	12	ALQDQLVLVAAK	13	36.5	7	5	23.5	24.1	956.6 (y9) 713.5 (y7)	956.3 - 956.6 713.2 - 713.5	13 12	45.6 20.0
Antithrombin-III (ENSG00000117601)	15	DDLTVSDAFHK	10	39.0	5	6	19.2	22.3	803.4 (y7) 704.3 (y6)	803.2 - 803.4 704.1 - 704.4	10 10	95.1 91.9
Apolipoprotein-A2 (ENSG00000158874)	3	SPELQAEAK	1	37.3	5	4	12.1	15	546.4 (y5) 659.4 (y6)	546.3 659.4	1 1	12 9
Apolipoprotein E (ENSG00000130203)	10	LGPLVEQGR	7	48.4	2	5	15.5	17.4	701.4 (y6) 588.3 (y5)	701.2 - 701.5 588.2 - 588.4	7 7	3.7 21.3
Beta-2-glycoprotein 1 (ENSG00000091583)	10	ATVVYQGER	9	36.0	4	4	12.4	15.4	652.3 (y5)	652.2 - 652.3	8	100
Ceruloplasmin (ENSG00000047457)	24	EYTDASFTNR	10	40.2	5	5	14.9	16.5	624.3 (y5) 695.3 (y6)	624.2 - 624.3 695.2 - 695.4	10 10	93 30.8
Clusterin (ENSG00000120885)	13	LFSDPITVTPVEVSR	37	29.8	10	9	28.5	26.3	1296.7 (y12)	1296.3 - 1297.5	36	27.9
Coagulation factor XIIIa heavy chain (ENSG00000131187)	8	VVGGLVALR	9	44.7	3	6	19.7	21.6	784.5 (y8) 685.4 (y7)	784.4 - 784.6 685.3 - 685.4	7 8	2.1 100
Complement C9 (ENSG00000113600)	10	AIEDYINEFSVR	13	37.9	5	6	28.3	31.3	1271.6(y10) 1027.5 (y8)	1271.2 - 1272.2 1027.2 - 1028.2	12 13	27.1 23
Complement factor H (ENSG00000000971)	33	SPDVINGSPISQK	23	34.4	7	6	16.3	17.7	830.4 (y8) 572.3 (y5)	830.1 - 831.2 572.2 - 572.5	17 23	93.6 30.3
Fibrinogen alpha chain (ENSG00000171560)	16	GSESGIFTNTK	15	40.3	4	7	14.7	17.3	610.3 (y5) 780.4 (y7) 867.5 (y8)	610.1 - 610.4 780.2 - 781.5 867.2 - 867.5	13 15 13	91.2 31.1 27.1
Fibrinogen beta chain (ENSG00000171564)	14	QGFGNVATNTDGK	4	46.7	6	6	13.5	14.6	706.3 (y7) 805.4 (y8)	706.2 - 706.5 805.3 - 805.4	4 3	89.3 20.3
Fibronectin (ENSG00000115414)	32	VTWAPPPSIDLTLNFLVR	8	28.13	4	6	38.2	30.4	977.5 (y8) 862.5 (y7)	977.54 862.9 - 863.2	1 3	4 5.3

Haptoglobin beta chain (ENSG00000197711)	7	VGYSVSGWGR	28	41	5	5	18.2	17.9	562.3 (y5) 661.3 (y6)	562.1 - 562.4 661.1 - 661.4	28 28	100 37.9
Histidine-rich glycoprotein (ENSG00000113905)	10	DSPVLIDFFEDTER	29	33.7	7	7	37.7	27.7	1171.5 (y9) 1058.4 (y8)	1171.2 - 1172.25 1058.03	29 29	41.6 94.3
Hemopexin (ENSG00000110169)	15	NFPSPVDAAFR	31	59.5	3	6	23.6	20.4	959.6 (y9) 775.3 (y7)	959.2 - 960.2 775.1 - 775.5	14 14	92.6 8.6
Heparin cofactor II (ENSG00000099937)	11	TLEAQLTPR	6	52.4	5	5	16.4	18	814.4 (y7) 685.4 (y6)	814.3 - 814.4 685.3 - 685.4	5 5	100 34.2
Histidine rich glycoprotein (ENSG00000113905)	10	DSPVLIDFFEDTER	29	33.7	7	7	37.7	27.7	1171.5 (y9) 1058.4 (y8)	1171.2 - 1172.2 1058.0 - 1059.0	29 29	41.6 94.3
Plasma retinol-binding protein precursor (ENSG00000138207)	5	YWGVASFLQK	27	47.5	5	6	35.1	24.1	849.5 (y8) 693.4 (y6)	849.2 - 849.5 693.1 - 693.5	25 26	100 49.4
Plasminogen (ENSG00000122194)	16	LSSPAVITDK	3	51.0	0	6	15.3	17.7	743.4 (y7) 830.5 (y8)	743.3 - 743.5 830.2 - 830.5	3 3	9.7 14
Prothrombin (ENSG00000180210)	16	ETAASLLQAGYK	15	42.4	6	6	20.2	20.6	879.5 (y8) 679.4 (y6)	879.2 - 879.5 679.1 - 679.4	15 15	55.3 72.8
Serum albumin (ENSG00000163631)	30	LVNEVTEFAK	98	49.2	4	6	19.3	19.9	937.4 (y8) 694.4 (y6)	937.1 - 938.2 694.1 - 694.6	95 93	96.6 21.9
Serum amyloid P-component (ENSG00000132703)	7	VGEYSLYIGR	12	53.2	5	6	21.3	20.9	1057.2 (y9) 871.5 (y7)	1057.4 - 1057.5 871.3 - 871.4	10 11	12.6 49.4
Transferrin (ENSG00000091513)	32	EDPQTFYYAVAVVK	37	29.2	8	6	20.3	25.3	1160.6(y10) 1288.7(y11)	1160.2 - 1160.9 1288.0 - 1289.2	31 17	17.6 4.2
Transthyretin (ENSG00000118271)	5	AADDTWEPFASGK	78	38.9	7	7	22.3	18.3	921.4 (y8) 606.4 (y6)	921.1 - 922.2 606.1 - 606.4	78 77	56.7 99.6
Vitamin D binding protein (ENSG00000145321)	15	THLPEVFLSK	14	34.8	4	6	19.7	23.9	819.5 (y7) 932.5 (y8)	819.2 - 819.5 932.2 - 932.6	14 14	99.4 59.3
Vitronectin (ENSG00000109072)	9	DVWGIEGPIDAAFTR	48	32.7	9	8	36.4	42.0	947.5 (y9) 890.5 (y8)	947.1 - 948.2 890.2 - 891.2	42 42	84.2 33.8
		FEDGVLDPDYPR	27	40.6	5	6	22.2	24.9	875.4 (y7) 1031.5 (y9)	875.1 - 875.5 1031.1 - 1031.8	26 26	100 7.8
Zinc alpha 2 glycoprotein (ENSG00000160862)	8	EIPAWVPFDPAAQITK	13	29.9	8	8	36.2	27.3	1087.7(y10) 728.4 (y7)	1087.2 - 1087.6 728.2 - 728.6	12 13	97.5 7.8

Table 28 MRMaid performance using horse serum transitions. Settings applied were: species horse, default RT setting, 8-24 aa length, 80%, 50%. * - denotes that the peptide was predicted by MRMaid, but there was insufficient data to predict the product ion (performed 18th May, 2009).

Horse protein target (Ensembl gene identifier)	No. of peptides found for target (green)	Precursor peptide sequence used by Quotient	No. times peptide in GAPP	Average TS for peptide	No. of predicted (b-ions, 1+ only)	No. of predicted (y-ions 1+ only)	RT real (55 min) (min)	RT predicted (min)	MS/MS transition real (<i>m/z</i>)	MS/MS transition predicted (<i>m/z</i>)a	No. of times product ion seen	Mean relative intensity for product ion observationsb
Afamin (ENSECAG00000022497)	3	IAPQLSTEELTFLGK	2	31.99	0 (one 2+)	6	24.7	26.4	1124.6 (y10)	1124.54 - 1124.62	2	12.5
Angiotensinogen (ENSECAG00000015711)	8 (12 total)	AAEVGMLLNFMGFR	3	23.66	5	7	28.8	29.7	1185.6 (y10)	1185.59 - 1185.84	2	53
Apo C2 (ENSECAG00000014224)	2	STAAVSTYAEILTDQFLSLLK	2	28.35	7 (one 2+)	7	31.3	33.1	1064.6 (y9)	1064.15 - 1065.1	3	44
Complement factor H (ENSECAG00000011256)	4	YCDMPVFENAR	2	41.22	0	5	22.1	19.5	832.4 (y7)	832.379 - 832.458	2	100
Fetuin B (ENSECAG00000004201)	2	LSDASVLEAALESLAK	1	32.04	5	6	28.5	25.9	931.5 (y9)	931.538	1	72
Fibrinogen alpha (ENSECAG00000010083)	5	TFPGEGLDGLFHR	1	40.17	1	7	24.4	24.3	1187.6 (y11)	1187.59	1	8
		VLLSDLLPADFK	3	34.81	3	6	27.8	28.4	1005.5 (y9)	1005.43 - 1005.58	3	22
Fibrinogen beta (ENSECAG00000010239)	8	HQLYIDETVNSNVPTNIR	2	18.99	4 (one 2+)	7 (one 2+)	21.4	24.1	600.4 (y5)	600.445	1	100
		DNENVVGEYSSELEK	1	29.3	1	4	22.9	19.1	1041.5 (y9)	1140.65	1	42
Hemopexin (ENSECAG00000016475)	6	NFIGPADAAGR	1	37.33	0	4	22.6	21.7	804.4 (y8)	804.462	1	100
Inter-alpha-trypsin inhibitor heavy chain H1 AKA Inter-alpha (globulin) inhibitor H1 (ENSECAG00000024792)	2	GSLVPASAANLQAAR	1	36.72	0	6	20.7	20.5	1069.6 (y11)	1069.57	1	100
Maltase-glucoamylase (ENSECAG00000009647)	1	FAGFPDLITR	1	42.66	0	5	25.4	24.2	714.4 (y6)	714.429	1	100
Parotid secretory protein (ENSECAG00000000297)	3 (5 total)	ILPTVDGSLGLK	3	40.2	0	6	22.4	23.4	889.5 (y9)	889.34 - 889.618	3	22.6
Prothrombin (ENSECAG00000011321)	2	TTDEDFFLFFDVK*	1	38.9	3 (one 2+)	6 (one 2+)	27	28	508.3 (y4)	smallest is y5		
Vitamin D-binding protein	2 (7 total)	THIPEVFLSK (red)	1	41.8	3	5 (one 2+)	20.8	23.4	932.5 (y8)	932.561	1	21

(ENSECAG00000010481)												
Serum albumin (ENSECAG00000010305)	15 (21 total)	LPGSENHLALALNR	2	24.47	0	6 (2,2+)	20.3	22.7	1021.6 (y9)	1021.72	1	37
		TVLGNFSAFVAK	4	37.18	2 (one 2+)	5 (one 2+)	24.5	25.6	940.5 (y9)	940.367 - 940.487	4	54.25
Serum Paraoxonase (ENSECAG00000000344)	2	EVEPVELPNCNFVK*	1	47.13	2	1	23	23.3	878.4 (y7)	y9 only		
Apolipoprotein A2 (ENSECAG00000010163)	5	TQEQLTPLVK*	2	22.68	3 (one 2+)	2	19.6	20.1	456.3 (y4)	smallest y6		
		IGNDLLNFFSHFIELK	2	23.8	3	6 (one 2+)	29.7	33.4	1167.6 (y8)	1167.78	1	24
Clusterin (ENSECAG00000007010)	1	ASSIMDELFQDR	1	45.83	3 (one 2+)	7 (one 2+)	25.9	22.8	1053.5 (y8)	1053.49	1	79
Complement factor B (ENSECAG00000011854)	3	YGLVTYATVPK	2	47.03	0	6	21.9	20.8	779.4 (y7)	779.457 - 779.5	2	100

5.5.2 Retention time prediction is accurate

Compared to empirically-derived RTs (in Table 28), MRMAid's RT predictions from sequence alone are encouragingly accurate. The retention times measured for nineteen peptides were compared to the RTs predicted using Krokhin and co-workers algorithm (Krokhin *et al.*, 2004), as implemented in MRMAid (Figure 46).

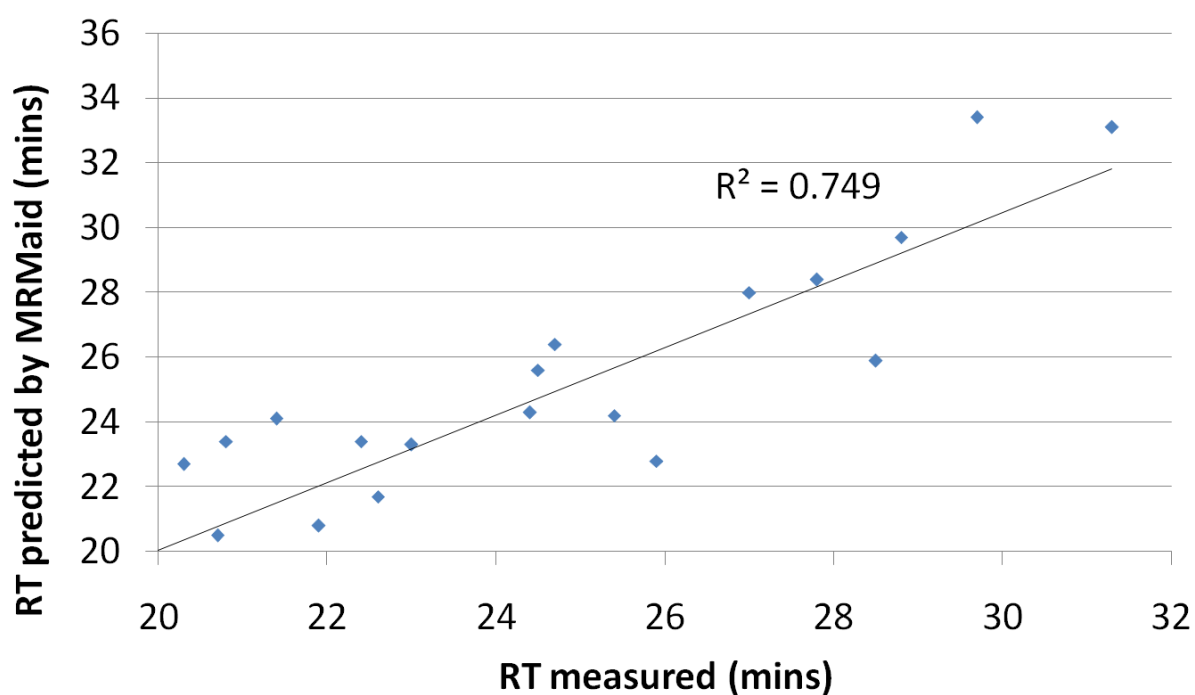



Figure 46 Observed versus predicted retention time for selected transitions shows that MRMAid's retention time prediction is accurate

The R^2 value was 0.75 (to two decimal places), indicating that the RT estimates suggested by MRMAid were very likely to be adequate for avoiding co-elution when planning a MRM experiments.

5.5.3 MR Maid has comprehensive documentation and user support

The author prepared resources to help users, and has made them available via the MR Maid website (see Figure 47 for a summary). These include: videos to introduce the aims of the system; demonstration 'walk-through' videos filmed using (a trial version of) Adobe Captivate™ 3; a glossary of terms; a user guide; and a link to the paper published in *Molecular Cellular Proteomics Journal*. Furthermore, interactive help 'bubbles' were integrated into the homepage, where users select the filters and enter details for their searches. The user guide including screenshots, glossary and journal paper can be found in Appendix V. In addition, a practical guide to using the MR Maid software has also been accepted for publication as a book chapter by Humana Press (2010), see Appendix V.



Summary of GAPP Database Content

- So far 8 proteins have been identified from 8 of your experiments
- 2964 proteins have been identified in 183 experiments using exclusively public data
- For a list of the experiments already processed by GAPP pipeline click here
- For further information go to the GAPP

MRMaid - GAPP's MRM transition design tool

Enter your protein target

Enter an accession number for your protein target (e.g. EN000000017427). This search uses EBI-PCRP

MS Instrument: MS:

Chromatographic conditions, go to this help bubble for descriptions

Species:

Tissue Type (e.g. Serum):

Select from the following options to increase the level of filtering performed on your results

- Do not accept internal cleavage sites
- Do not accept H or G in the peptide
- Do not accept M or C in the peptide
- Accept only peptides containing P
- Do not accept peptides containing F
- Exclude peptides with P3 (e.g. WFAK)
- Exclude peptides with P2 (e.g. WFAK)
- Do not accept peptides with E or G in the tryptic peptide
- Exclude minimum peptide length

MRMaid help area

Videos:

- part1: Come fly with us! [\(VIEW\)](#)
- Welcome to Cranfield University! Meet the MRMaid team and see how MRMaid was developed.
- part2: Final boarding call! [\(VIEW\)](#)
- Find out about the challenges we faced when developing MRMaid.

Tutorials:

- An introduction to Multiple Reaction Monitoring (MRM) and MRMaid [\(VIEW\)](#)
- What can MRMaid do for me? Why was MRMaid developed in the first place? This short presentation introduces the MRM approach for quantitative proteomics for those less familiar, and explains why MRMaid aims to solve one of the big challenges faced by researchers designing quantitative proteomics experiments.
- Getting started with MRMaid [\(VIEW\)](#)
- See MRMaid in action with this new video. This short video includes a demonstration of how to submit a query to MRMaid, an introduction into how to interpret the results and it shows you how to navigate between peptide and product ion information.
- Designing multiple transitions with MRMaid [\(VIEW\)](#)
- This video shows you how to download your results from MRMaid and explains how you can use the results to design a monitoring experiment for lots of different protein targets.

Glossary:


- Download the glossary pdf here
- For more information on what the terms mean that are used in MRMaid, check out this glossary.

Guide:

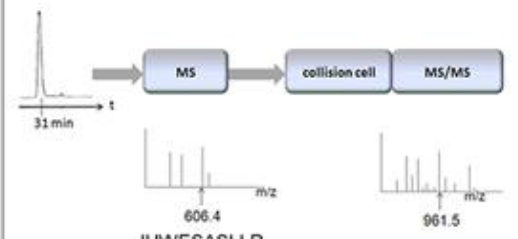
- Download the user guide pdf here
- This is a detailed user guide to lead you through the process of designing transitions.

Paper:

- View the MRMaid manuscript at MCP
- This is a paper in press at Molecular and Cellular Proteomics Journal. It was published online on November 15th, 2009



MRM: multiple reaction monitoring



SRM transition_{complementC3} = 606.4 / 961.5 at RT = 31 min

Dictionary of terms used in MRMaid

average TS

The peptide candidates selected by MRMaid each have an average TS (transition score) value associated with them. The average is calculated over all of the times the peptide was identified in GAPP database (for the protein of interest). For each time the peptide sequence is assigned to the protein, a slightly different TS is observed. This is because fragment ion information is included in the TS calculation, and the ions generated are likely to be slightly different on each occasion the peptide is seen. The mean average is taken across all these individual assignments. See 'TS' for details of the calculation.

b-ion


A molecular ion produced when the peptide bond in a peptide breaks during fragmentation in MS/MS. It is the N-terminal part of the peptide. The b-ion has a number according to the breakage position compared to the N-terminus. MRMaid highlights b-ions in blue in the product ion results view. See also 'y-ion' and 'daughter ion'

c

See 'peptide coverage'

chromatographic condition

The type of reverse phase chromatography used. MRMaid estimates retention time (RT) for each peptide using a published model. To refine the RT estimate for a particular set-up users can select from a series of options and setup-specific correction factors will be applied to the calculation. See 'retention time'



Molecular & Cellular PROTEOMICS

Originally published in *Proteomics* at doi:10.1002/prob.2009231867208 on November 15, 2009

Molecular & Cellular Proteomics 8: 694-705, 2009
© 2009 by The American Society for Biochemistry and Molecular Biology, Inc.

Research

MRMaid, the Web-based Tool for Designing Multiple Reaction Monitoring (MRM) Transitions*

Jennifer A. Mead¹, Luca Basso², Vanessa Ottome², Claire Burtone², Richard G. Kay³, Kathryn S. Lilley³, Nicholas J. Bantle² and Conrad Bessant^{2,4}

From the ¹Bioinformatics Group, Cranfield Health, Building 63, Cranfield University, College Road, Cranfield, Bedfordshire MK43 0AL, United Kingdom; ²Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 3EQ, United Kingdom; and ³Cambridge Bioscience Ltd., Wellesley Road, Wellesley, Cambridge CB7 2PP, United Kingdom

Figure 47 Summary of the MRMaid user documentation and help resources taken as screenshots from www.mrmaid.info (26th August, 2009)

5.6 Discussion

5.6.1 MRMAid can design transitions with multiple product ions

MRMAid is essentially an SRM design tool which can be used for MRM design by combining the results of several SRMs in a single spreadsheet. As explained above, a protein accession number is entered via the interface and candidates are predicted. The results may be downloaded, which include both product and precursor ion data, then following this, the next protein to be targeted is entered. Candidates are generated by MRMAid and downloaded, as before. This process is repeated until all the spreadsheet data has been captured. Finally, using TS, number of observations and RT, the candidates may be ordered for experimental validation in MS/MS, before purchasing synthetic surrogates or designing an expression construct (see Figure 48).

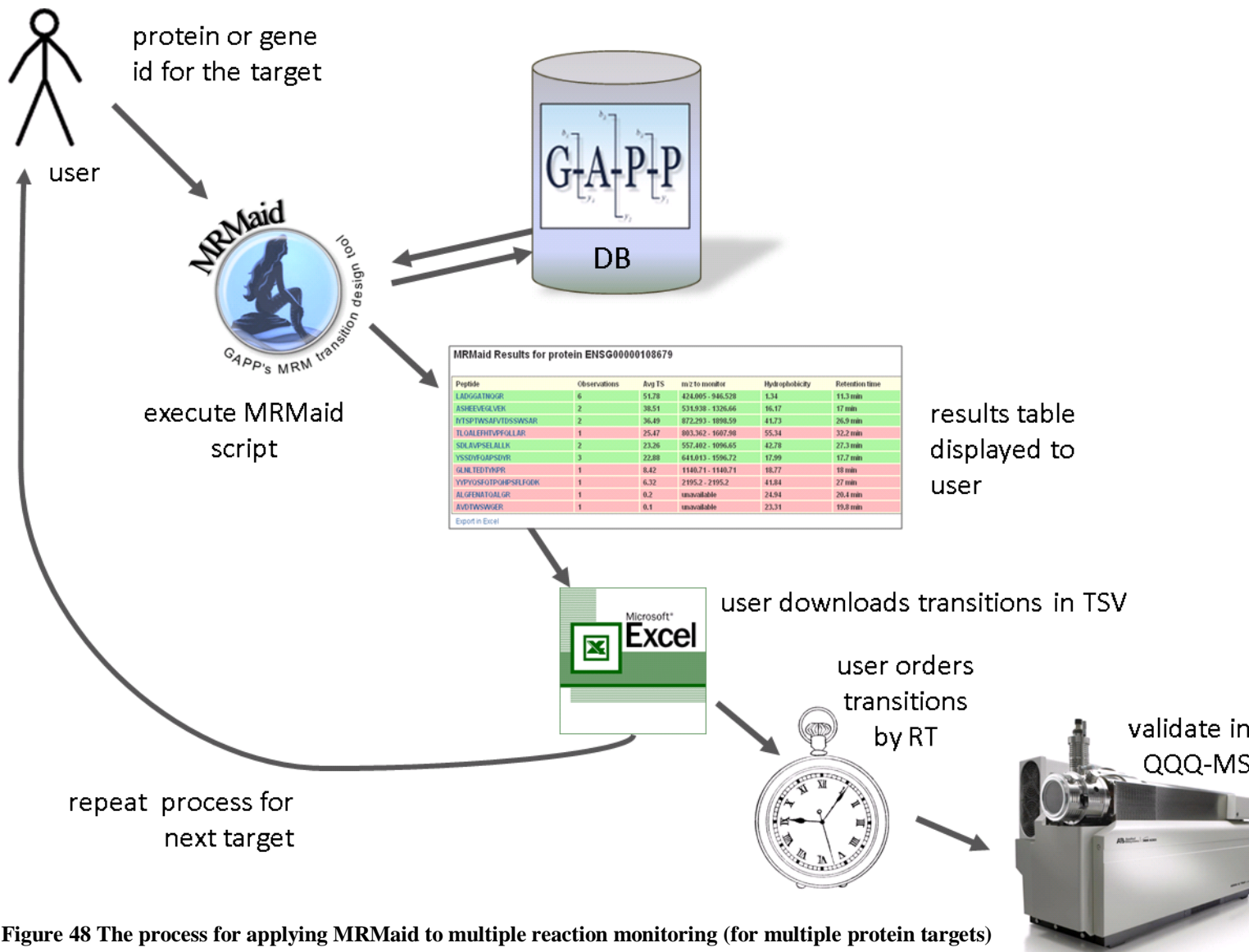


Figure 48 The process for applying MRMaid to multiple reaction monitoring (for multiple protein targets)

Recommendations and tips to assist users in optimal candidate selection are provided in the interactive help documentation.

5.6.2 MRMAid can design multiple transitions for targeting proteins in complex samples

MRMAid is particularly suitable for designing transitions with multiple product ions. A shortlist of multiple peptide candidates is provided for each target, so two or three of these can be selected for validation. For each of these, a list of many different product ions is provided, each with metrics to indicate their reproducibility and reliability. At this stage users may select several of the best product ions. In this way, MRMAid supports multiple product ion selection for each peptide. This allows monitoring of a target even in very complex samples: by having several ions to act as signposts for the protein in MS/MS, it can make it possible to confirm its presence in spite of high levels of noise.

Moreover, by allowing users to apply a sample type filter at time of search, such as serum, transition design for complex samples is further supported, because this strategy increases the likelihood that a candidate will be selected that has been proven to work even amongst associated sample noise.

5.7 Conclusions

MRMAid is a modular, web-based tool built around the GAPP DB framework, providing a web-based solution for fast and reliable transition design. Candidate transitions are ranked based on a novel transition scoring system, and users may

refine the results by selecting optional stringency criteria. Comparison with published transitions showed that MRMAid successfully predicted the peptide and product ion pairs in the majority of cases with appropriate retention time estimates.

It is web-based with an intuitive user interface and does not require computer expertise to use - avoiding download and complicated processes for setup locally. It also supports diverse MS instruments and RP chromatography conditions, so has flexibility to be useful to a large number of users.

MRMAid is a tool for designing MRM transitions and is intended to support the proteomics community by exploiting public data resources and prior knowledge of the MRM technique. MRMAid eliminates the need for time-consuming preliminary studies by delivering ranked candidate transitions based on an existing repository of experimental data - meaning that far fewer transitions need to be validated before a suitable candidate is found. The software is freely available as an executable application on the web at www.mrmaid.info and is a major deliverable for this EngD.

As with any automated workflow, no human judgment could be applied to interpret the results; only widely applicable rules could be used. Despite this, the MRMAid results presented demonstrate that accurate peptide-product ion transition predictions can be made when MS/MS data is available for querying. Estimation of

RT is also of high enough accuracy to be useful for ordering transitions and avoiding co-elution.

As the data content of the Genome Annotating Proteomic Pipeline repository increases, the coverage and reliability of MRMAid are set to increase further. In conclusion, MRMAid represents an effective first step towards the future of integrated software applications for the design of quantitative proteomic experiments.

In the wider context for proteomics research, MRMAid is a fitting example of a tool that contributes to transformations in knowledge (McNally, 2008). Specialised disciplinary skills and expertise from lab-based MRM practitioners were built into MRMAid in a process that now enables novices and other relatively unskilled personnel to perform what previously was a ground-breaking, expert procedure – namely designing targeted assays for several targets. This philosophy is an accepted way that modern science, including proteomics, progresses in the ‘knowledge economy’.

5.8 Suggested developments for MRMAid releases in the future

5.8.1 Transition candidate ranking using a star rating

The MRMAid transition ranking system could be improved by linking MRMAid with MRMAid database, a repository of published transitions (see Chapter 6). The

principle is to rank the transitions by a star rating or as 'gold', 'silver' or 'bronze', depending on the level of MS evidence for the transition candidate in question, by MRMAid calling the MRMAid-DB during processing. The transition candidates could be ordered into three categories, for example:

- Gold transitions: when a published, validated transition from a journal publication is available in MRMAid-DB for the target protein
- Silver: when the transition candidate is supported by MS/MS data in GAPP database
- Bronze: when the transition candidate is supported by a theoretical prediction only, i.e. when MS/MS evidence or a published transition is absent

The benefit of this would be to make it even clearer to the user which candidate is the best, increasing the likelihood of reproducibility.

5.8.2 Batch mode for submitting protein targets

MRMAid can predict transitions for only one protein target at a time. Newer additions to the field, such as MaRiMba (Sherwood *et al.*, 2009), support batch submission of proteins, so that transitions for several proteins may be determined simultaneously, saving the user time inputting and interpreting the outputs of searches. Support for batch submission of targets would therefore improve the MRMAid program.

5.8.3 Protein sequence as input

Collaborators at Quotient Bioresearch and the University of Cambridge have suggested that MRMAid would be greatly improved if it could accept protein sequences as inputs, not just protein accession numbers. In this way, users can skip the time needed to find the protein ID and there is less chance of an erroneous ID

being used. As in Skyline (Prakash *et al.*, 2009) - a recent addition to the array of transition design tools - protein or peptide sequences can be input directly by the user and filters are applied to decide which peptides are most suitable, given 'rules' of transition design (similar to the MRMAid approach); no accession numbers are needed.

5.8.4 Interspecies mode for predicting transitions

Generally, MS/MS data for some species, such as human and yeast, is more readily available compared to other species used in biology research, such as the rat, horse, dog or hamster, for example. Having a facility in MRMAid to predict transitions for less well-supported species may prove useful in the future. This would require extrapolation from the data that is available. In genomics, for example, the interspecies approach can be applied to microarray analysis, to analyse CHO cell lines using mouse arrays, for example (Ernst *et al.*, 2006). However, it remains to be seen how this approach may be applied for data-driven, cross-species transition design. Although cross-species identification for MS using PMF has been demonstrated (Lester *et al.*, 2002), the problem of leveraging data from one species to design transitions to target a protein in another has yet to be addressed *at all* in the literature. If MRMAid could be modified to offer this service, it would be unique, and from communications with collaborators, such as researchers at Quotient BioResearch who design MRM assays to target horse proteins, it would add value to the field, potentially helping many research groups.

Having protein sequences as inputs *per se* (as described in the previous future work item) would be compatible with the implementation of interspecies transition prediction in MRMAid. By searching for any transition candidates that have suitable spectra for a given peptide sequence (regardless of which species database the target protein pertains to) would mean candidates could be searched for in an unbiased way – with species not constraining the search.

5.8.5 PSI-compliance of MRMAid

A standard format for storing and sharing transition data, TraML, is almost complete. Ideally, MRMAid should be modified to make it possible to export the transition candidates it predicts in a fashion compliant with the new standards. If this is not possible, because MRMAid does not have the minimal data required for the final agreed version, then MRMAid should at least be modified to export the transitions in a format compatible with MS instruments, so that the candidates can be directly programmed into the instrument ready for validation and/or quantitation; this way researcher avoid having to convert the list from MRMAid manually.

5.8.6 Porting data to MRMAid via GAPP

In essence, MRMAid is a data-mining algorithm for retrieving transitions from a database of spectra. The predictions it can make are only as good as the data quality and content of the repository it mines. Ideally, therefore, MRMAid should be adapted to work as a universal ‘front-end’ program that can interrogate a suitable ‘back-end’ MS/MS data repository. For example, data from other public

repositories, such as PRIDE, GPMDB or Tranche, could be ported to the MR Maid server, or indeed a configurable, generic version of MR Maid could be coded to sit on top of these public databases. GAPP DB is too limited and is less likely to continue to grow at a fast rate as the other offerings, such as PRIDE.

Moreover, the generic, configurable version of MR Maid could be implemented as a commercial software product for mining customers' in-house MS/MS data repositories, and also mine public datasets as well (at the same time), if desired. This would satisfy the unmet need for companies, such as OBT (mentioned earlier), who wish to have a completely private facility for predicting transitions with MR Maid.

MRMaid-DB: a repository for published MRM transitions

*“Once an SRM assay for a protein is established,
it becomes universally useful and exportable”*

From Picotti *et al.*, 2008

6.1 Summary

The thesis so far has focused on repositories for disseminating proteomic MS data and identifications. In this chapter, a new type of repository is developed, namely a public database for storage and dissemination of published and experimentally validated transitions. It is designed for researchers seeking to quantify proteins using the increasingly popular technique of SRM, providing a structured system for published transitions, which otherwise would remain in publications in disparate forms, being difficult to systematically search. The database is unique compared to the only other offering in this space (MRMAtlas), because it directly couples transitions to the research paper from which they came, and permits users to submit their own transitions as they publish them.

Together, MRMAid (described in the previous chapter) and MRMAid-DB (described here) provide a two-pronged approach to address the lack of computational resources for the SRM practitioner. For example, when there are no published transitions, users may use MRMAid to design new ones, and when there are validated transitions in the literature, they can interrogate MRMAid-DB to retrieve them.

6.2 Introduction

The key to selected reaction monitoring (SRM) is finding the best peptide-to-product ion transitions to monitor. The MRMAid database (MRMAid-DB), presented in this chapter, is a new online database for capturing SRM transitions from published research papers to save practitioners time when searching for transitions that have been previously validated. It contains all the information needed to reproduce the transitions, such as information on the sample matrix, HPLC, and MS instrumentation used, and also includes details of the manuscript of origin. Transitions are submitted using simple web-based data entry forms, meaning researchers have a simple way to increase access to their transitions, and in turn, may increase the citations for their research papers.

To use a crude analogy, MRMAid-DB may be compared to a second-hand shop, where researchers can bring their own SRMs and search for new ones as they need them, only the 'goods' are all free. However, MRMAid-DB does not contain a haphazard mixture of good, average and poor quality transitions, but instead only contains high quality, experimentally validated transitions; each one having been published in peer-reviewed journal articles. Each SRM entry is checked manually upon entry, and the product ion type and mass are cross-checked against the peptide sequence to ensure accuracy. In MRMAid-DB, the previous 'owners' of each transition are available -as journal titles and author names- and users can quickly determine if a transition will fit in and suit their own lab because all protocol, sample

handling and mass spectrometry details are available. MRMAid-DB's 'shop window' is based on the Biomart framework, using colour schemes and 'look and feel' consistent with the MRMAid/ GAPP family of software.

6.3 An introduction to MRMAid-DB

The key to performing successful SRM studies is finding the best peptide-to-product ion transitions to monitor. To find the best transitions, researchers can perform empirical validation themselves in the laboratory, or exploit the new tools for transition design that have emerged during this EngD (Walsh *et al.*, 2009, Martin *et al.*, 2008, Lange *et al.*, 2008, Prakash *et al.*, 2009, Mead *et al.*, 2009, Sherwood *et al.*, 2009). The most reliable transitions, however, are those that have already been experimentally validated and peer-reviewed by others. Yet the problem with using these is the lack of a common reporting format, and the absence of a central resource for published transitions, meaning researchers must scour the literature manually taking time and effort. The aim in producing MRMAid-DB, therefore, was to solve this problem by providing a freely accessible collection of the majority of available empirically-confirmed transitions, described in proteomics literature to date. Furthermore, it is a scalable, structured database system to support the continued growth in generation of SRM data in the community.

6.3.1 Existing repositories for SRM transition data

The only other SRM transition data resource on the web at the time of writing was MRM Atlas (Picotti *et al.*, 2008). In the MRM Atlas paper, this group reinforced the idea that storing validated transitions is a scientifically worthwhile effort, because

they demonstrated that transitions may be reproduced across different QQQ-MS instruments (Picotti *et al.*, 2008) with minimal 'tweaking', such as collision energy adjustment. MRM Atlas is built on the SBEAMS framework (Marzolf *et al.*, 2006), and contains ready-to-use peptide-to-product ion transitions for approximately 1,500 *S. cerevisiae* proteins (equivalent to 21% of the yeast proteome) (Picotti *et al.*, 2008). Information stored includes protein sequence; precursor charge; Q1 *m/z*; Q3 *m/z*; intensity; ion type; collision energy; hydrophobicity; observed retention time; number of peptide observations; annotator (name of individual/lab or 'best'); and a hyperlink to a 'consensus' spectrum (representative pattern of product ions for the peptide by averaging many observations). MRM Atlas is searchable via PeptideAtlas, or may be queried via its own interface (see (Picotti *et al.*, 2008) for an example). Furthermore, using a new data visualisation option, SRM assays in MRM Atlas can be accessed by clicking on protein targets in metabolic pathway diagrams.

MRMaid-DB is distinct from MRM Atlas in several ways: firstly, it contains transitions and the infrastructure to support transitions for multiple species: currently there are optimised transitions for monitoring proteins from horse, yeast, cow, mouse and human. It is also unique in providing a quick and easy method of uploading transition data online, and includes paper manuscript details, so users can extract all the transitions from a given article in one search.

At the time of writing, plans were presented for a new resource: the PeptideAtlas Transitions Resource (PATR)¹⁰⁰. It is not available yet, but demonstrates the topical and fast-moving nature of this field of research.

6.4 Method and implementation

6.4.1 MRMaid-DB Data Content

A typical transition has the following core information: a protein target, peptide sequence (including any PTMs), precursor m/z ; product ion m/z ; and observed peptide retention time. MRMaid-DB stores this core information, but in addition, it stores detailed auxiliary information for reproducing the transition accurately.

The database scope was set by the author by capturing user requirements for SRM assays from lab-based practitioners. Once a list of data-fields had been formulated, the items were cross-checked with nine research papers (Gerber *et al.*, 2003, Kuhn *et al.*, 2004, Zhang *et al.*, 2004, Cox *et al.*, 2005, Anderson and Hunter, 2006, Stahl-Zeng *et al.*, 2007, Kay *et al.*, 2007, McKay *et al.*, 2007, Keshishian *et al.*, 2007) that included SRM assays to see if the desired data items were usually reported in peer-reviewed papers (without needing to contact the authors). For the majority of the items on the list the data was routinely available, although some decoding and knowledge of the field was required to extract the information, such as decoding the descriptions of the RP-HPLC protocols, for example.

¹⁰⁰ Deutsch *et al.*, ASMS conference 2009, Philadelphia USA, poster presentation

Detailed information is captured so that users can easily identify the most suitable transition for their particular workflows; data-items include sample processing, biological matrix, and RP-HPLC-MS setup parameters. All of these factors can affect the reliability of detection of the precursor/product ions, so by making these available users have the best possible chance of reproducing the selected assay(s) in their own labs. Moreover, if the transitions were validated for a given sample type, it follows that they should be reproducible for the same sample type if variables are kept constant. Indeed, by specifying the details of biological matrix in MRMAid-DB, it ensures that the orders of magnitude and interference can be accounted for when users' samples, settings, and platform match those of the validated transition(s) in the database.

6.4.2 The MRMAid-DB transitions database schema

The established data fields were encoded into a non-redundant database schema, designed by the author. A database does not need to be a faithful representation of reality (see Figure 49 for the steps used), because it is not a simulation or model. Rather, it is a more pragmatic solution, whereby only the features are included that are absolutely necessary for the purpose of the system to be fulfilled. As an analogy, a database is a play, which depicts a real-life story, but only picks out events that are necessary for the audience to get the message to be conveyed. The entities in MRMAid-DB and the relationships between them have, therefore, been chosen with just two things in mind: does the user need this information?, and is the data available in journal papers?

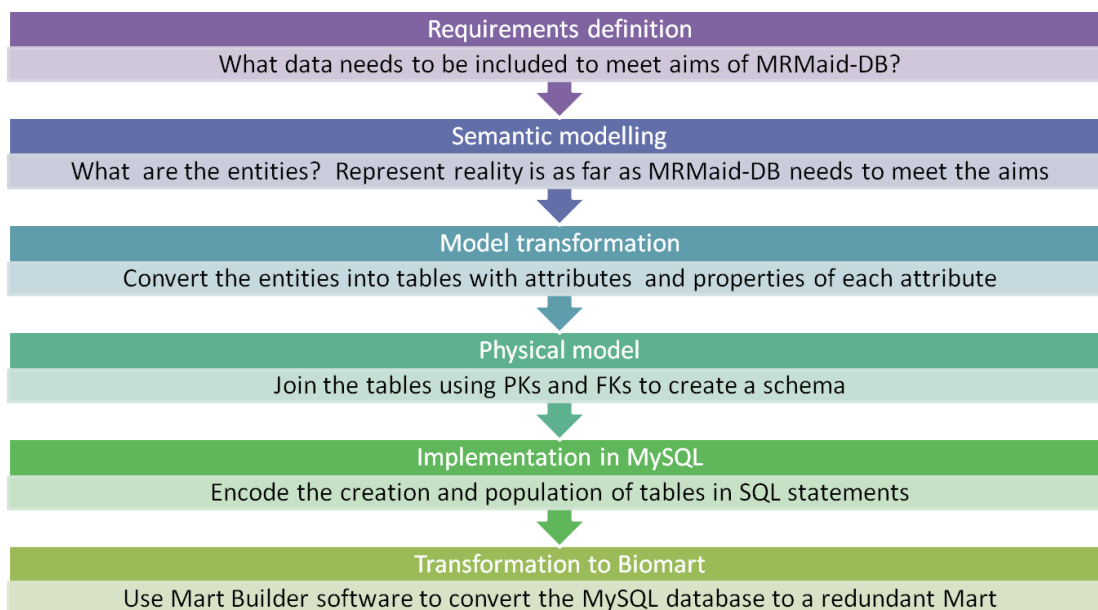


Figure 49 The methodology used to create the MRmaid-DB transitions database. PK is primary key, FK is foreign key.

The database schema was created in Fabforce.net DB Designer 4 for Windows, a free database design suite. Using this package, the tables and relationships were entered and the schema could be exported to a ‘create database’ SQL script automatically.

The MRM transition schema includes 16 tables, ten of which are look-up tables (Figure 50). One-to-many relationships exist between the tables; for example, in the look-up table ‘journal’, one journal contains many individual papers, so is a one-to-many relationship. Likewise, each paper in the ‘paper’ table can contain many transitions so paper to transition table is also one-to-many.

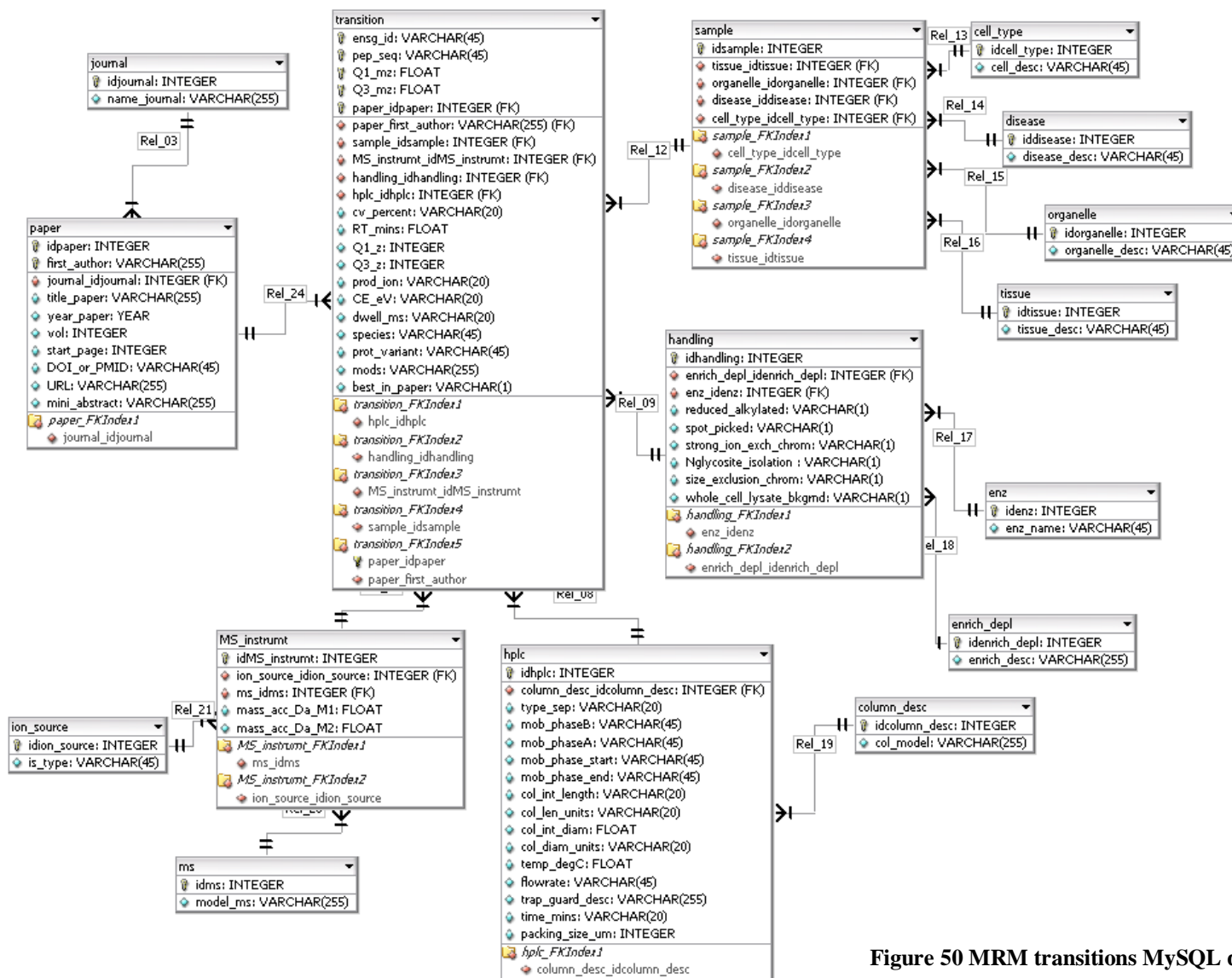


Figure 50 MRM transitions MySQL database schema

The tables were initially populated in batch with data taken manually from the nine papers (those mentioned earlier) by creating a comma delimited file of the data (see Appendix VI) and uploading it table by table. Tables were populated using PHPMyAdmin, a web-based software that is part of the Easy PHP (2.0.0.0) installation bundle.

In some cases, papers used protein IDs such as Swissprot or IPI for the protein targets, for which there were multiple Ensembl gene ID equivalents; for example, IPI00303963 (Complement C2 precursor) which is cross-referenced (by PICR, accessed 10th June, 2009) to ENSG00000166278, ENSG00000204364 and ENSG00000206372. To cope with this, all Ensembl gene IDs were included in the data entry for the first instance of MRMAid-DB.

6.4.3 MRMAid-DB is based on the Biomart framework which offers flexible and federated querying

The type of database front-end to use for populating and querying the database was selected next. The database was implemented in MySQL, which is a widely used database framework that is compatible with several database 'skins'. The author investigated several alternatives that were free and compatible with MySQL databases¹⁰¹. These were:

¹⁰¹ Intermine (www.intermine.org) was considered but it is compatible with postgresSQL, not MySQL.

1. DadaBik¹⁰², a free PHP-based front-end that is customisable and can support searching, inserting and updating of database records. DadaBik was tested (see Figure 51) but was ruled out because multiple primary keys could not be supported.
2. Xataface¹⁰³, a flexible front-end that automatically generates web forms and search menus. It offers customised features and functionality via configuration files, templates, and plug-ins. Xataface was ruled out because it was excessively rich in features and flexibility, a simpler solution was preferred.
3. Biomart¹⁰⁴, this interface is familiar to biologists, because existing databases in biology research use it. It offers very powerful, flexible querying and is supported by its creators, who offer implementation support by email.
4. Code a solution from scratch using PHP, as was done for GAPP database's website. This would require coding all possible queries in SQL statements manually into web-forms, taking more time but having the possibility to make the site totally bespoke. This option was ruled out, because time was better spent on novel aspects of the system, not coding an interface from scratch - a routine task in IT.

¹⁰² www.dadabik.org

¹⁰³ <http://xataface.com>

¹⁰⁴ www.biomart.org

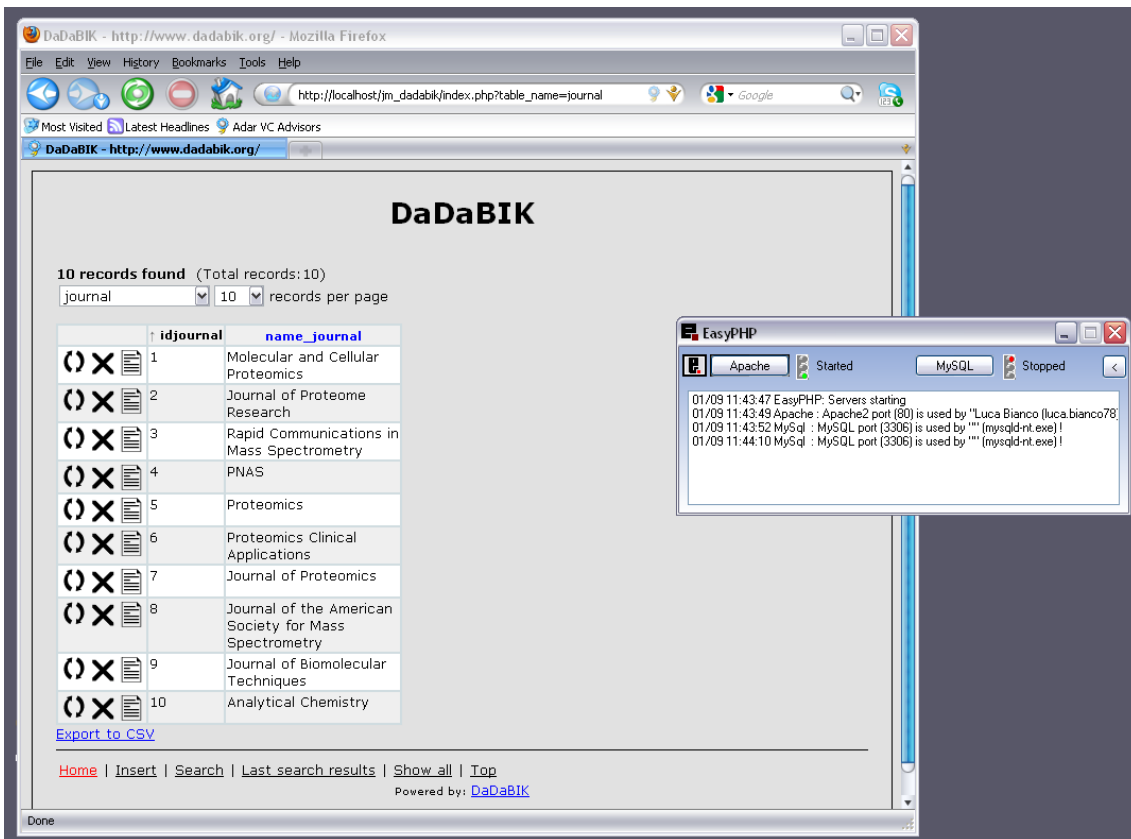


Figure 51 Dadabik was tested as a possible solution for web-based querying and data entry to the transition database

Biomart was chosen, because it offered superior implementation support and scope to federate queries with existing EBI Biomarts, advantages which were valuable for this project and are explained later. Furthermore, the database in this chapter was developed at the end of the EngD, so a solution had to be selected that could be implemented quickly, given the time constraints. Biomart offered the quickest route to have a working prototype compared to the other options.

Biomart does however, have limitations; for example, it is less flexible than coding from scratch, because the interface is configurable only within the parameters offered by the system. Moreover, the author could not incorporate interactive help pop-ups in the interface or other functionality. It was also not possible to 'drill down' into results, where one query is run, the results displayed, and then these results queried further by the user.

6.4.4 Implementing an instance of Biomart

Pre-requisites for running Biomart version 0.7 on the web were:

- Martj-0.7: notably this includes Martbuilder and Marteditor, all the configuration files, such as header.tt, and other files to manage the appearance of the Mart interface.
- LAMP (Linux Apache MySQL and PHP) configuration: both the Biomart *per se*, and the other pages of the MRMAid-DB website were web-based, so must be set up to run on a server.
- Biomart Perl: available via Perl package manager (PPM)
- Correct configuration of XML files, such as the registry files

Once the MRmaid-DB schema was created in MySQL, the database was transformed into a Biomart database, using the Martbuilder (Figure 52). A Biomart database is a database with greater redundancy, so queries can be executed quickly via the finished Mart query interface.

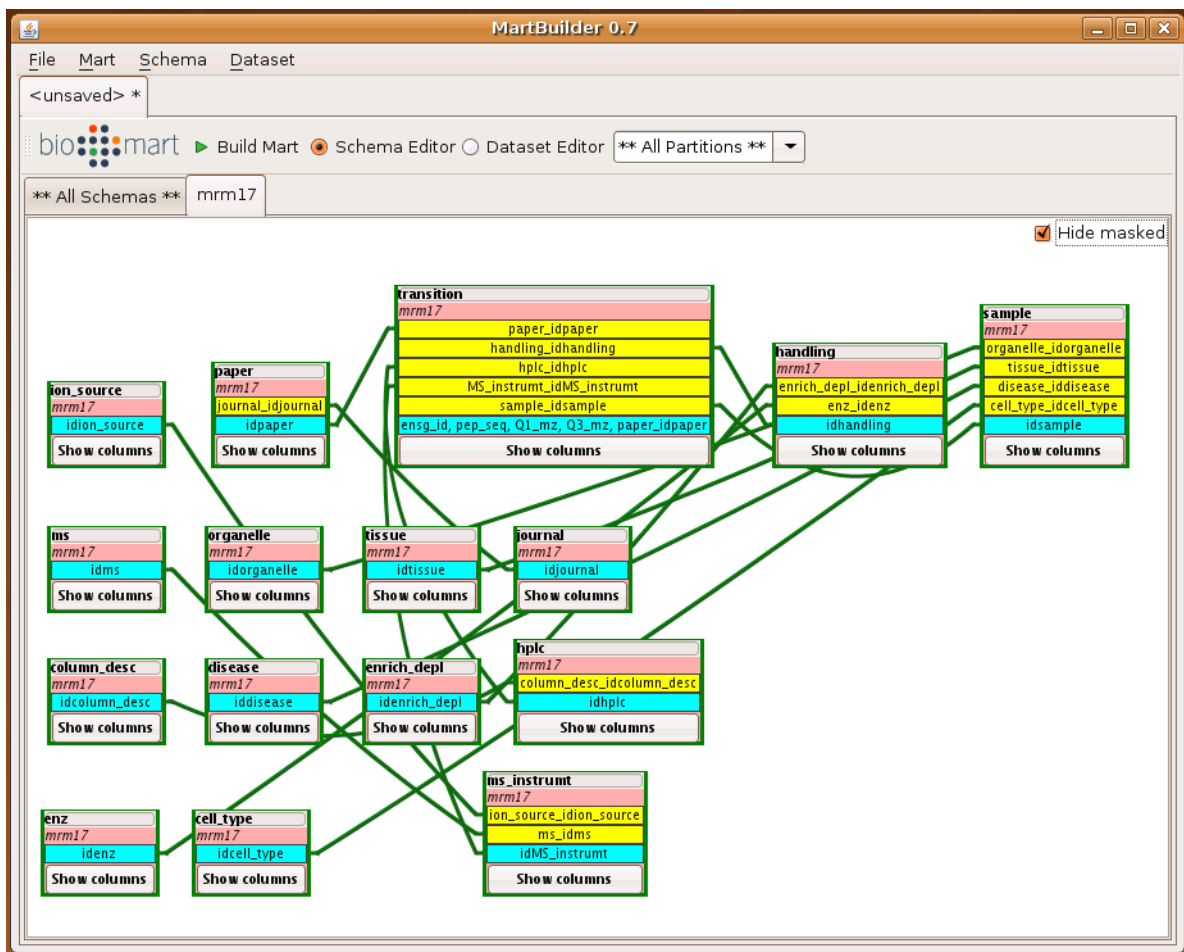


Figure 52 The MRM transition database (a MySQL database) shown in Martbuilder. The MySQL database was transformed to a Biomart using Martbuilder software.

Martbuilder transformed the MySQL transitions database (16 tables) into a large redundant table ('main') table, and one other ('dm') table that stores the display

parameters for specifying the interface appearance (which is populated by exporting information from Marteditor, this is explained later).

The Biomart 'main' table is based on the transition table, because it is the central table in the original MySQL schema (Figure 53). The new mart version of the database is exported as a 'create SQL' script (Figure 53, window inset), hence the Biomart is also encoded as SQL, so is also a MySQL database, like to original, but has only two tables instead of 16.

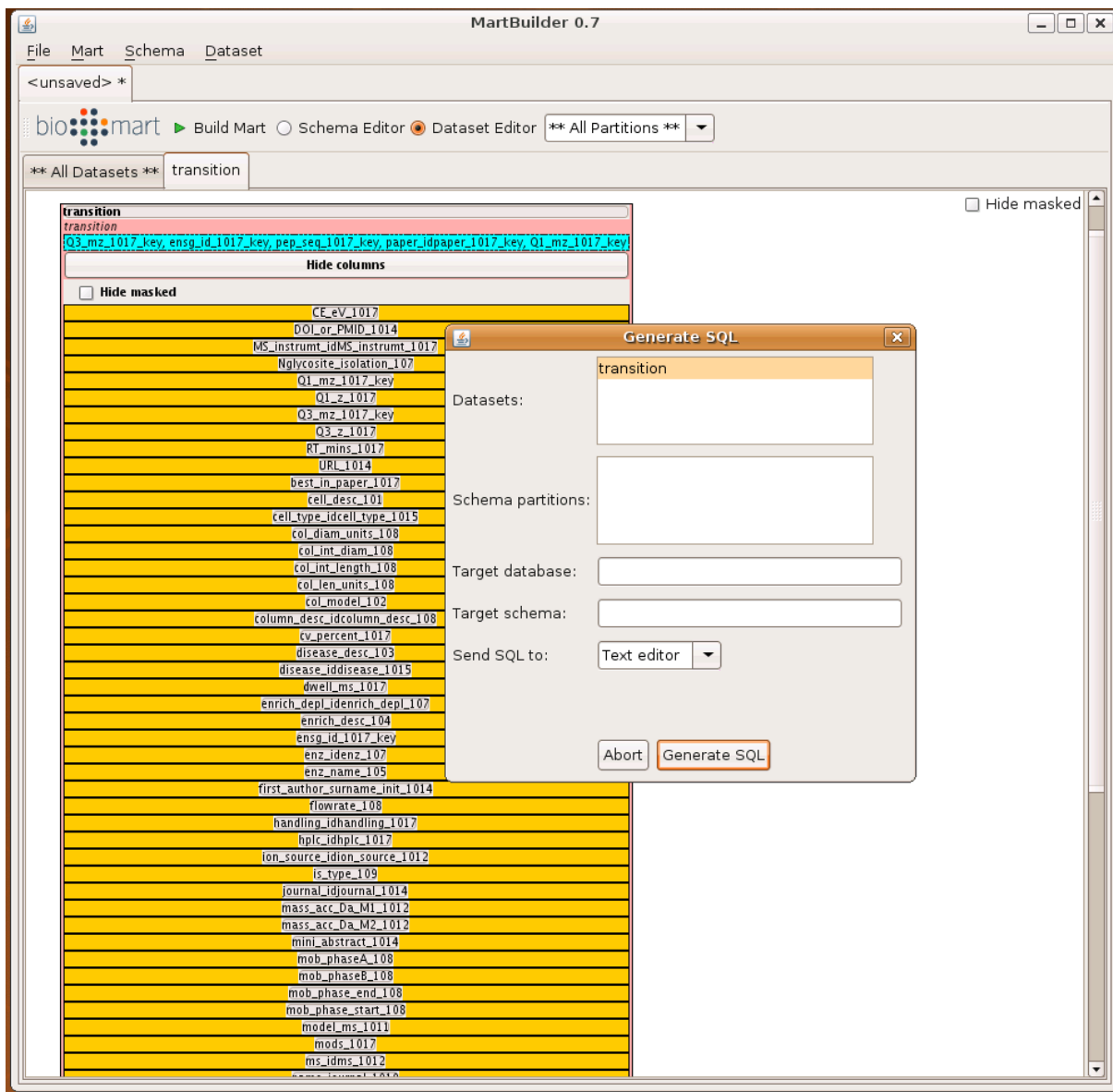


Figure 53 Martbuilder transforms the MySQL transitions database (16 tables) into a large redundant table this 'main' table is based on the transition table because it is the central table in the original MySQL schema.

Next, the Biomart MySQL database was imported into Marteditor (Figure 54).

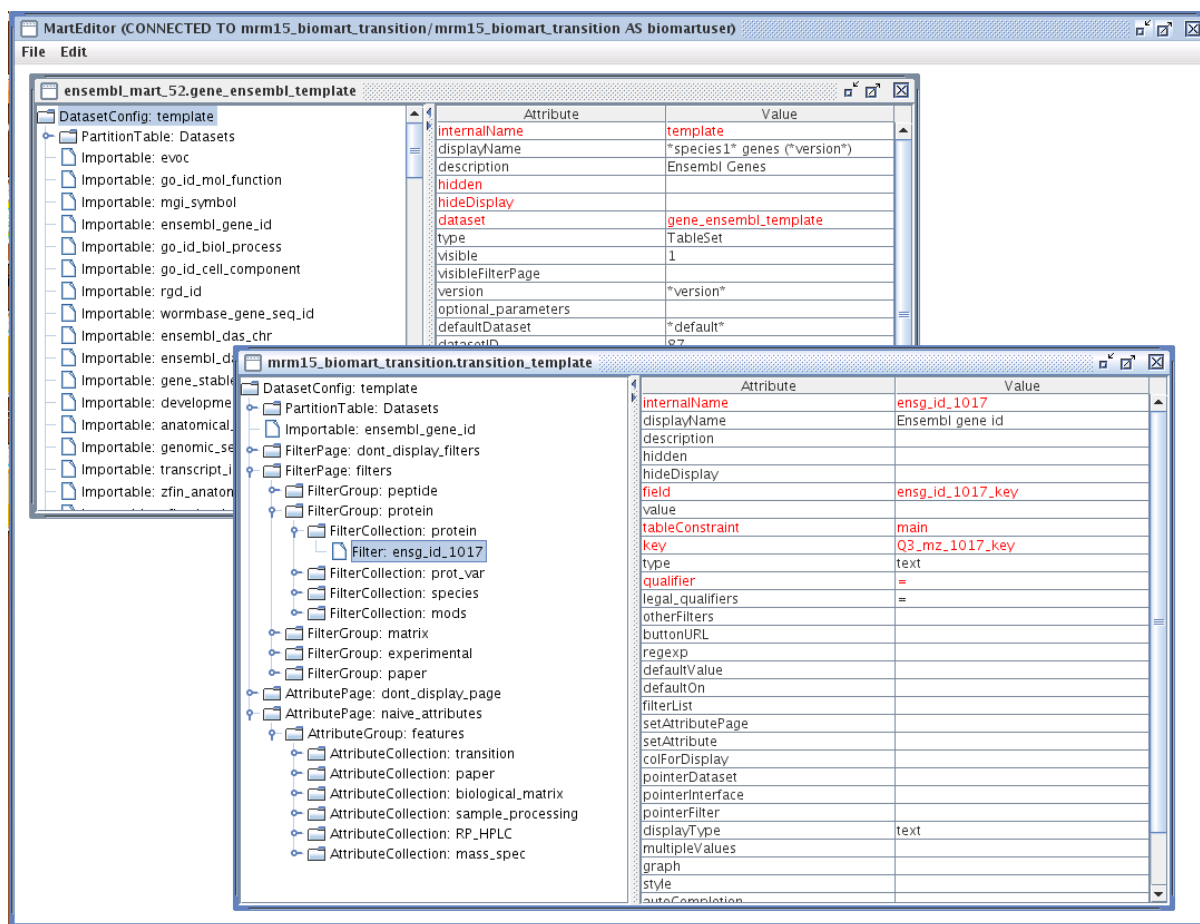


Figure 54 Marteditor software was used to tune the Biomart interface parameters, including the attributes and filters displayed

In Marteditor, one can configure the interface displayed to the user, specifying the attributes and filters, for example; this system of 'filters' and 'attributes' is now explained briefly.

The Biomart query interface is a database front-end that has standard features: it has *filters* - criteria on which the results are filtered/constrained (a protein ID, for example) and *attributes* - the data items that the user wishes to have returned from the search (the peptide, *m/z* values and collision energy, for example). Biomart

allows any combination of filters and attributes to be selected simultaneously, offering the maximum level of query flexibility to users.

Theoretically, any of the data-fields in the Biomart could be applied as filters or attributes in the interface, if this was specified in Marteditor. The author has chosen only the filters and attributes deemed pertinent to the needs of users of MR Maid-DB (shown in Figure 55). The assumptions for this selection were based on discussions between the author and expert MRM practitioners. There is a worked example of using filters and attributes in Case Study 1 in the results section of this chapter.

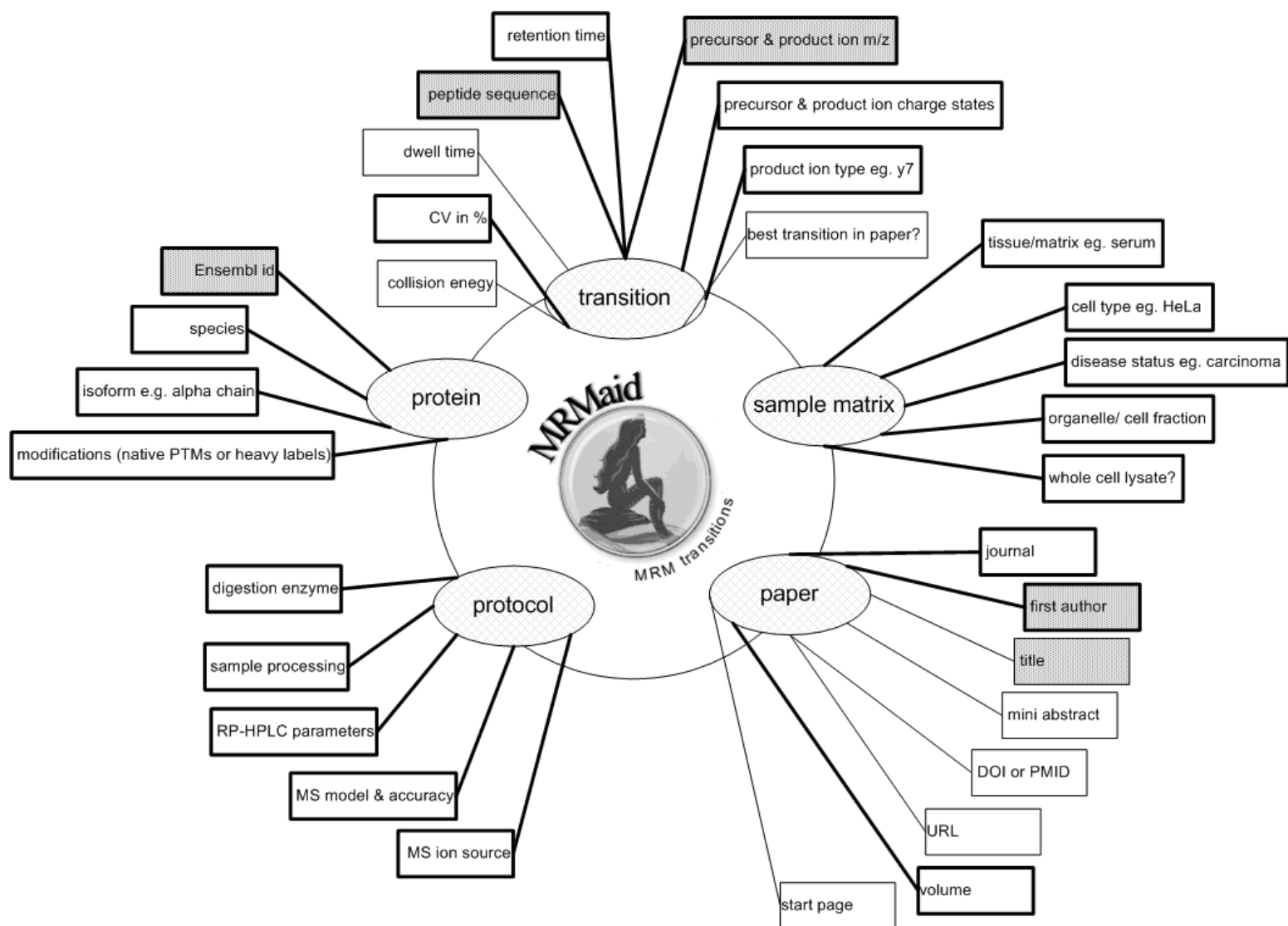


Figure 55 Filters and attributes in the MRMAid-DB transitions Biomart. The ellipses indicate the five main data categories in the Biomart query interface. The boxes show the data-fields available in the schema. Shaded boxes indicate data-fields that are compulsory when submitting transition data. The boxes with a heavy line are the data-fields that can be applied as filters when searching the Biomart, the other boxes are just attributes

The parameters entered into Marteditor for the filters and attributes must all refer to the new Biomart database column names and table names, not those of the original database and must be compatible with the other configuration files. Once these are set, the settings are 'exported', which means they are added automatically by the Marteditor program to the 'dm' table of the transition Biomart MySQL database.

6.4.5 Federating the MRM transitions Biomart with Ensembl Biomart

The MRM transition Biomart was federated with the Ensembl Biomart, an existing database provided by the EBI, by joining the Ensembl gene ID column. Using this setup, the two Biomarts can be queried simultaneously. The advantage of this is that the most up-to-date descriptions for targets may be retrieved from the EBI servers, rather than retrieving descriptions or names of targets from a local version of Ensembl, which may become out-of-date. There are five logical sections to the transition Biomart: transition, protein, protocol, matrix, and paper (Figure 55). Each section has several data-fields, as per the original schema. Only linking fields, such as foreign keys are not displayed to users.

To get the Ensembl Biomart options to display correctly, the settings had to be specified in the registry XML file. The complete set of Biomart configuration files are in Appendix V. Configuring Biomart was not trivial, since the written documentation did not have explicit instructions on how to do it. The author sought advice and practical help for setting up the system, including the federated querying, directly from the Biomart developers.

6.4.6 Submitting transition data to MRmaid-DB

The website allows any registered user to submit SRM assays, which are then manually curated by the administrator (for now, the author) and inserted into the database (Figure 56).

As soon as the submission is approved, the data becomes immediately available via the public interface by execution of two PHP programs: the first moves data from the temporary user submission database, to the MySQL database, and the second transforms the updated MySQL database to a Biomart MySQL database, compatible with the query interface of the website (Figure 56). These two PHP programs were coded by Luca Bianco.

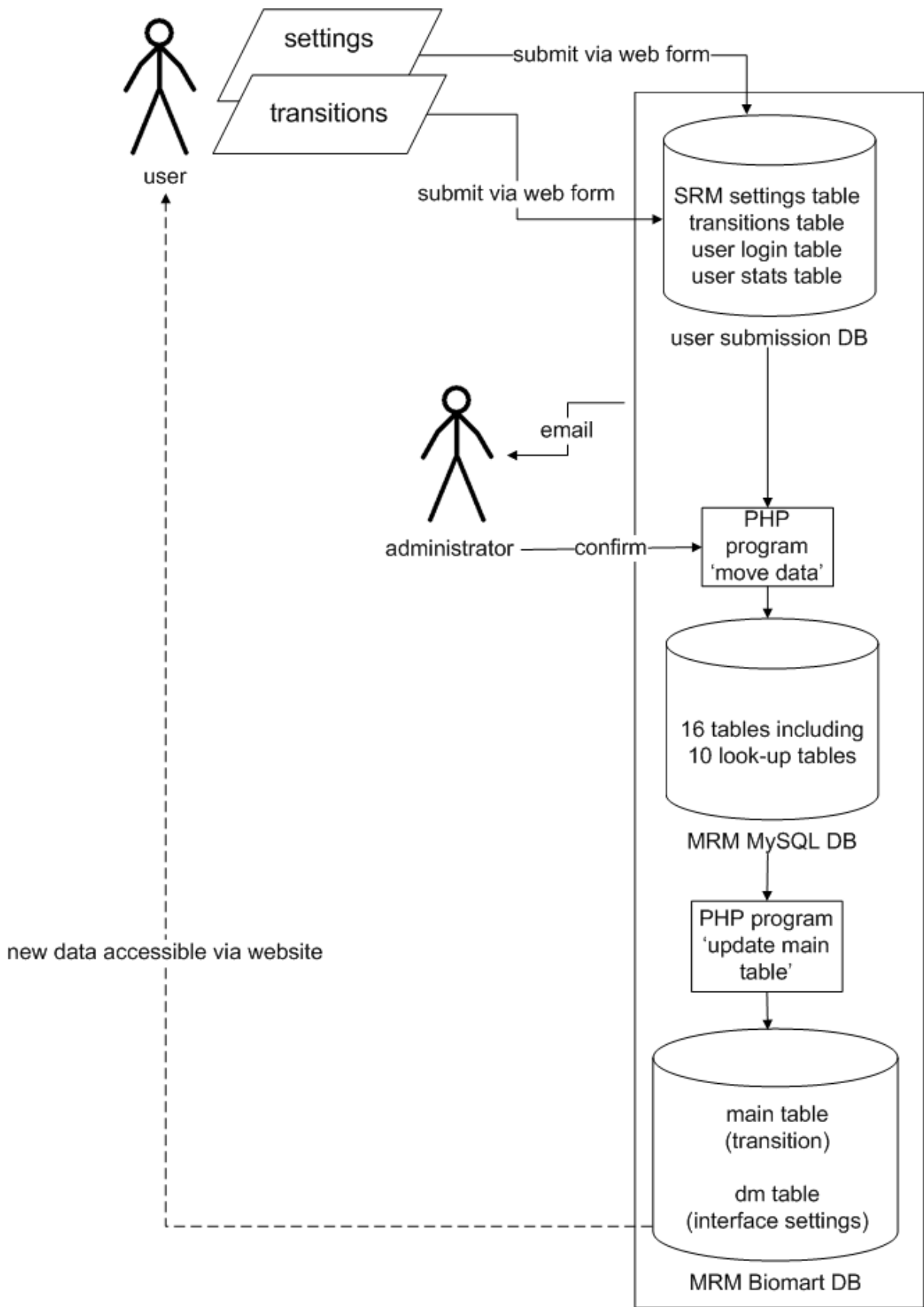


Figure 56 The process of data submission to MRmaid-DB

The data submission process (and hence the temporary user submission database) were designed by the author on the basis of how practitioners perform and publish MRM studies. Multiple transitions can be performed given a single experimental protocol, and multiple transitions may be published for single protocol/workflow; thus, to save users time when submitting transitions this idea was applied. For example, users must only submit biological matrix, protocol, and manuscript information once, save the information as a list of settings, and give the list a name. These settings can then be applied when submitting transition data on subsequent visits to the site (see Case Study 3 in the results section). The temporary user submission database (Figure 57) stores the submissions in this logical format as an intermediate stage before entry into MRMAid-DB proper. The data entry system was conceptualised and designed by the author, but to save time for the author (who was extracting and submitting transitions from papers for the first release of MRMAid-DB at the time), Luca Bianco encoded the items into a MySQL schema, ready for use on the MRMAid-DB website.

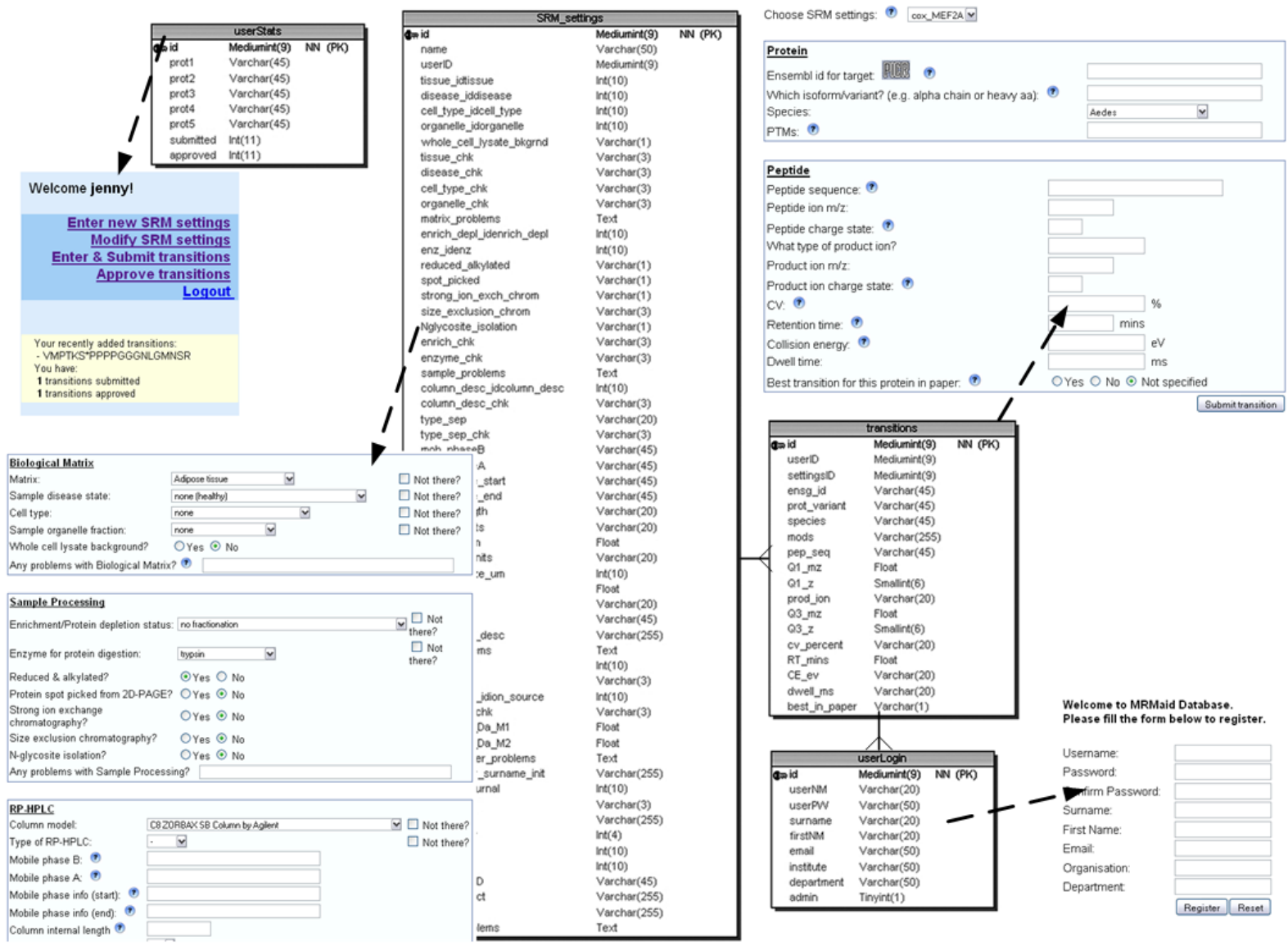


Figure 57 Data submission database tables for the MRMAid-DB website. Dotted arrows show the website interfaces for the corresponding MySQL tables. The solid lines indicate links between database tables in the schema.

To test the submission process, selected transitions from a recent paper (Barton *et al.*, 2009) were entered by Chris Barton (the author of that paper), via the completed MRmaid-DB web interface. His submission tested the process shown in Figure 56, and helped to ensure that the data entry was intuitive for an external user. His comments were applied to improve the interface. Screen shots of the whole data submission process are available in the user guide in Appendix VI.

6.4.7 Exporting data from MRmaid-DB

When a query is executed, the attributes selected by the user are presented in a single results table, which can be exported in HTML, CSV, TSV or XLS formats. The results can be refined to show distinct rows of data by selecting a check box in the results display panel. This export functionality is standard for Biomart.

6.4.8 Designing the MRmaid-DB website look and feel

MRmaid-DB is a comprehensive website, not just a Biomart query interface. Biomart has been applied by other groups. One offering (see <http://paramecium.cgm.cnrs-gif.fr>) was used as an example of how to create a coherent website around a Biomart query interface. There is a homepage, help area and information about the creators for users to browse.

6.5 Results

6.5.1 Examples of MRMAid-DB use cases

The following case studies give detailed walk-throughs for the most typical procedures performed using MRMAid-DB. The author has also produced a user guide, which is available in Appendix VI, which also includes example workflows.

Case study 1: Retrieving a list of transitions for a protein target

In this scenario, a list of validated transitions is retrieved for a specific protein target, namely human apolipoprotein A-II (ENSG00000158874). The filter must be selected first, this is the field on which data retrieval is to be restricted; in this case, it is the protein identifier. To begin, the user clicks 'Browse Mart', chooses the 'MRM transition database', clicks 'filters' on the left toolbar, and expands the 'Filter by protein' section. The Ensembl ID of interest is then pasted into the 'Ensembl ID for target' text box - the adjacent box must also be checked to indicate this filter is active. The Ensembl ID now appears on the left toolbar.

Next, the attributes are selected, which are the data-fields to be retrieved for this protein. The user clicks 'Attributes' in the left toolbar to see the available lists of attributes. For SRM transitions, useful information includes the protein target (in this case the name and ID), peptide, precursor m/z (Q1), product ion m/z (Q3), peptide retention time, collision energy and dwell time. These features are now selected by the user as attributes, by checking the relevant boxes.

The protein name (description), which is desired for the output list, is not available in the attribute form shown (for transition Biomart); it must instead be selected from the Ensembl Biomart. To do this, the user clicks the second 'Dataset' link in the left toolbar and chooses in the drop-down menu the dataset for Homo sapiens (NCBI36); this is because the target in this example is human, so the dataset must be selected depending on the species in the query. To confirm the user's selection, the database name is now shown in the toolbar. Next, the user clicks 'Attributes' for the Ensembl Biomart, and expands 'Gene'. Ensembl Gene ID and Ensembl Transcript ID are checked by default. The user unchecks these and selects instead description (and/or Ensembl Protein ID, not used in this example); again the selections now appear in the left toolbar.

The filters and attributes are now in place, so to retrieve the transitions the user clicks the results button above the toolbar, which executes the search. The results show the transitions retrieved (screenshot in Figure 58) with the table including data from both the transition and Ensembl Biomarts. To export these results, the user selects the download option from the drop-down menu above the table and hits 'Go'.

Welcome to the MRmaid DB BioMart

MRmaid Database
High quality, published SRM transitions for quantifying your proteins of interest

Cranfield UNIVERSITY

HOME BROWSE MART SUBMIT DATA GO TO GAPP HELP ABOUT

New Count Results URL XML Perl Help

Export all results to TSV Unique results only Go

Email notification to

View rows as Unique results only

Ensembl gene id	Peptide	Q1 m/z	Q3 m/z	Retention time (mins)	Collision energy (eV)	Dwell time (ms)	Description
ENSG00000158874	EPC*VESLSVQYFQTVTDYGK	1175.6	1221.5	64.02	40	30	Apolipoprotein A-II Precursor (Apo-AII) (ApoA-II) [Contains Apolipoprotein A-II(1-76)] [Source:UniProtKB/Swiss-Prot,Acc:P02652]
ENSG00000158874	EPC*VESLSVQYFQTVTDYGK	1175.6	1436.6	64.02	40	30	Apolipoprotein A-II Precursor (Apo-AII) (ApoA-II) [Contains Apolipoprotein A-II(1-76)] [Source:UniProtKB/Swiss-Prot,Acc:P02652]
ENSG00000158874	SPELQAEAK	486.8	546.4	12.1	-1	-1	Apolipoprotein A-II Precursor (Apo-AII) (ApoA-II) [Contains Apolipoprotein A-II(1-76)] [Source:UniProtKB/Swiss-Prot,Acc:P02652]
ENSG00000158874	SPELQAEAK	486.8	659.4	12.1	-1	-1	Apolipoprotein A-II Precursor (Apo-AII) (ApoA-II) [Contains Apolipoprotein A-II(1-76)] [Source:UniProtKB/Swiss-Prot,Acc:P02652]

Figure 58 Screenshot of results from a federated query using MRmaid-DB for apolipoprotein A-II. The first seven columns are from the MRM transition Biomart, and the final column is from the Ensembl Homo Sapiens (NCBI36) Biomart database. The asterisk (*) in the peptide sequence indicates a modification on the previous residue. '-1' is entered for numerical fields when the data is not available in the manuscript ('null' if a non-numerical field). The tool bar on the left displays the filters and attributes selected for the search; the first dataset refers to the MRM transition Biomart, and the second dataset refers to the Ensembl Biomart. For both peptides shown, there are two possible transitions. Simultaneous monitoring of multiple transitions that identify the same peptide can increase selectivity of the SRM assay; MRmaid-DB can store multiple transitions pertaining to the same peptide and protein, for this purpose.

Case study 2: Retrieving all the transitions from a particular manuscript

A frequently cited paper is Anderson and Hunter's investigation, where human plasma proteins were monitored using SRM (Anderson and Hunter, 2006). This case study shows how to retrieve all the validated transitions from this particular article.

To begin, the user selects as filters for the transition database: first author 'Anderson NL'; journal 'Molecular Cellular Proteomics'; and volume '5' (volume is optional in this case). Next the user selects the attributes: Ensembl gene ID; peptide; precursor m/z (Q1); product ion m/z (Q3); and retention time. As explained in case 1, selecting 'description' in the Ensembl Biomart gene attributes is the way to retrieve the target name, if required. To count the number of transitions for this paper, the user clicks the 'count' button above the left toolbar. This shows that there are 119 transitions that meet the filter search criteria (displayed below the count button). To retrieve the transition data for these, the user clicks the 'results' button. A list of the transitions and protein descriptions is displayed, and can be exported.

Case study 3: Submitting transition data to MRMAid-DB

For users to start submitting transition data from their own manuscripts, they must first register and log in via the 'Submit data' area of the website. Once logged in, a submission status bar appears on the left to indicate the number of transitions submitted and approved for this specific user. To begin submission, information on the experimental set-up must be entered and saved (as 'SRM settings'). Clickable help bubbles (blue question marks) are positioned next to some of the data-fields in the data entry form: these give a brief explanation of the type of data required, with specific examples. During the data entry process, if the data required is not available in the drop-down menu options then users must check the 'Not there?' box and write in their desired entry in the text box at the end of the section (labelled 'Any problems...?'). In this case, the administrator will enter the user's request into the

relevant lookup table of the schema, making it available for future submissions. This feature is required because comprehensive controlled vocabularies were not available for all aspects of the MRMAid database, such as HPLC column models and mobile phase compositions.

Once the form is complete, the settings can be saved. If there any data entry issues, these are highlighted in red when the user attempts to submit the form. The data fields that are flagged can be amended and the form submitted again, without losing the settings already entered.

To submit individual transitions, the user selects the link on the left. The settings that were entered and saved earlier are now selected by the user in the drop-down menu. Once selected, the transition data entry form is filled in by the user. The protein target must have an Ensembl accession number; if the user has another type of accession number it must first be converted to Ensembl. To convert the accession number, the user can click on the PICR icon; this takes him/her to a tool that cross-references database accession numbers and can convert lists of various IDs in batch mode (Côté *et al.*, 2007).

After completing the form, the user clicks 'Submit transition' to send the data to the database administrator for approval. Once approved, the transitions enter the Biomart immediately and are available to all users.

6.6 Discussion

There is currently no officially sanctioned standard for reporting SRM transitions in research papers and repositories. HUPO-PSI and Institute of Systems Biology are currently coordinating efforts to develop a standard XML-based format for storage and exchange of SRM transitions, called transitions markup language (TraML). The first draft was posted on the HUPO-PSI website¹⁰⁵ on May 2nd, 2009. The overlap between the proposed TraML standard format (as in August 2009) and MRmaid-DB has been examined (Table 29).

¹⁰⁵ www.psidev.info

Table 29 A comparison of the data captured by TraML versus MRmaid-DB. The comparison shown here is based on the TraML early draft (Version 0.2.0.0) released on May 13, 2009.

	Transitions Markup Language (TraML)	MRmaid Database
Data-fields specified as compulsory	TraML, version: TraML format version used CV, URI: uniform resource identifier for the controlled vocabulary CV, fullName: name of the controlled vocabulary CV, id: identifier for the controlled vocabulary CV, version: version of the controlled vocabulary contact, id: contact person for referencing publication, id: identifier for the publication instrument, id: identifier for the MS instrument software, id: identifier for referencing the software software, version: version of the software used to generate transitions protein, id: identifier for the protein target protein, name: name of the protein target cvParam, accession: accession number for the controlled vocabulary cvParam, cvRef: reference to the controlled vocabulary parameters cvParam, name: name of the controlled vocabulary peptide, id: identifier for the peptide peptide, modifiedSequence and unmodifiedSequence: peptide sequence with and without modified amino acids compound, id: identifier for the compound to be monitored by SRM precursor, mz: Q1 m/z product, mz: Q3 m/z prediction, softwareRef: reference to the software for transition prediction prediction, transitionSource: how the transition predictions were made (e.g. using a consensus spectrum search) configuration, instrumentRef: reference to instrument configuration for validation or optimisation of transitions validation, transitionSource: how the transition(s) was validated	Ensembl gene ID (for the target) Peptide sequence Q1 m/z Q3 m/z First author (of the paper) Title (of the paper)
Transition targets	Different biomolecules	Peptides only (and validated in peer-reviewed papers)
Signal intensity data	Signal intensity, intensity rank, and relative intensities	No intensity information
Redundancy	Some redundancy if all elements are filled in, such as Q1 m/z which can be specified in Element <transition> and Element <precursor>	Redundancy is avoided by storing the SRM settings only once for each type of setup, and reusing this for the individual transition submissions.
Sequence information	Whole protein sequences and peptides	Protein IDs and peptide sequences
Product ion data	Ion series and ordinal separately: 'y' and '6', respectively.	Product ion type as 'y7', for example.
PTMs and heavy peptides	PTMs or heavy residues are written as square brackets in the modified peptide sequence.	PTMs are specified by an asterisk in the sequence and a text field with the name of the modification(s). Heavy residues are described in the isoform / variant data-field

Publication details	Published articles from which the transitions are derived can be captured, as Pubmed IDs, for example.	More detail on the paper, including the Pubmed ID or DOI, as well as journal name, title, mini abstract, volume, start page, etc.
Contact details	All contact names of people who produced the transitions	First author of the paper article
Transition design / analysis software	Name and version of the software used to design the transitions. Also the 'transition source' (how the transitions were selected)	No software or transition selection information
Peptide retention time	Predicted or actual RT data	Observed RT only
Protein name/ description	Protein description is in TraML (in Element <protein>)	Transition data Biomart database is federated with the EBI's Ensembl Biomart for several species (currently human, mouse, dog, cow, horse, and yeast). The description/name of the target ID must be retrieved from the Ensembl Biomart.
Coefficient of variance	-	Captures coefficient of variance in % for each transition from the published paper manuscript.
Ranking transitions	Transitions ranked in recommended order by the experimentalist	For each protein in each paper the transition is indicated as the best (y) or not the best (n) or unknown (0), if not specified in the research paper.

Although TraML and MRMAid are very different (the former is a data standard, whereas the latter is a database) it is worth comparing their data models, because it would make sense to support TraML import/export within MRMAid when TraML becomes stable. Currently, MRMAid-DB export is in a simple spreadsheet format, although results can be extracted as XML via the Biomart API.

TraML contains the specifications for representing SRM transitions for monitoring different types of compounds via mass spectrometry, not just proteins and peptides, plus signal intensity information, and experiment/sample information encoded using ontological terms. This contrasts with MRMAid-DB which, being dedicated specifically to proteins / peptides, captures only a subset of the data included in the TraML specification. Furthermore, MRMAid-DB also takes a pragmatic approach to data entry, allowing free text entry for fields for which ontologies have yet to be developed, such as for specific RP-HPLC conditions.

At the moment there is no relevant MIAPE (The Minimum Information About a Proteomics Experiment) module for this area of proteomics, so TraML and MRMAid-DB enforce their own rules regarding compulsory data fields. Again, MRMAid-DB currently takes the more pragmatic approach, requiring only a subset of the data considered compulsory in TraML (Table 29).

6.7 Conclusions

MRMaid-DB is a scalable compendium of high quality, validated transitions available at www.mрмаid-db.info. It provides a freely accessible online framework to store and disseminate transitions as they become available in the literature. This has the potential to save users valuable time when designing SRM experiments, and to increase citations for those authors submitting their transitions to the database.

Along with TraML and MRM Atlas, MRMaid-DB acts as a useful template for researchers looking for guidance on what to include in their SRM publications in the future. Indeed, in the same way that prototype software applications act as a useful tool for communication between users and developers, these new developments will hopefully stimulate researchers to discuss and decide which aspects of experimental information are most important for reproducing SRM assays in practice.

6.8 Future work

6.8.1 Data content of MRMaid-DB

For the purposes of this thesis, the novel work was development of the data management system, not database population. At the time of release, MRMaid-DB contained only 272 individual SRM transitions, which is sufficient for a prototype system but does not reflect all of the published transitions in the public domain at that time. MRMAAtlas, the only other system for storing MRM transitions, is limited to yeast at present, but despite this has thousands of transitions already. Immediate

future work, therefore, should be population of MRMAid-DB with newly published transitions.

6.8.2 MRMAid-DB should become compatible with standards

MRMAid-DB could be made TraML-compatible (hence PSI-compliant), once the standards are finalised. This was not possible at the time of writing this EngD, because the standard was not ready for release, however, as demonstrated in Table 29 there are already data-fields in common. For example, a parser program could be produced to allow export of entries from MRMAid-DB in TraML format for the data-fields available.

6.8.3 Automated quality control at point of data entry

Each SRM entry is checked manually by the author to verify that product ion type and mass are correct using both the research article and also web-based m/z calculators to ensure fidelity of the data entered. This is a process which could, at least in part, be automated. In MRMAid, for example, there is already a function to calculate the b- and y-ion complement of a peptide sequence. Using this code, the mass of the peptide entered could be automatically checked against the m/z values (and ion type) entered by the user into the form.

Conclusion

This project contributes to the body of knowledge in proteomic bioinformatics on several levels. Firstly the field of 'proteomic bioinformatics' was characterised in terms of its commercial presence (Chapter 2), and as an array of publicly available repositories and software resources (Chapter 3). There is no research available in the public domain regarding the management of proteomic bioinformatics, so it is believed that the work presented in Chapter 2 adds significant value to the field. The immediate future plan for the material is to re-format it into a teaching case for the MBA course, for example via the European Case Clearing House¹⁰⁶. The business history could also be written up separately as an article for *Enterprise & Society: The International Journal of Business History*. The review of the latest developments in public repositories, Chapter 3, offers a further notable contribution to the field, because the public repositories are constantly changing in functionality and content. A review in a single document is thus valuable to both lab-based proteomics practitioners, who wish to upload their datasets or mine existing ones, and for proteomic bioinformaticians, wishing to have an overview of existing resources to avoid repetition of effort and to design tools to plug gaps in the resources currently available.

In Part II, the long standing question of which decoy database should be used to reduce FPs in automated identification workflows was answered. A systematic investigation of diverse decoy database designs was carried out using GAPP,

¹⁰⁶ A not-for-profit organisation that provides a collection of case studies and papers for teaching purposes for management training.

revealing that the recommended decoy database design is peptide level reverse, searched independently from the target database. This work was a 'brute force' empirical approach and offers a practical way to reduce the risk of pursuing false protein leads in biological/pharmaceutical research. Naturally, by taking the recommended approach, which stringently filters out FPs, it may happen that true identifications are also filtered out thus not providing the full picture of the proteins present in a sample. In this scenario, the investigator can be sure that what is identified was present, but will not know what is absent from the list - this is favourable in a pharmaceutical setting, for example, when researchers must be certain that the protein was present. The work also has potential to affect reporting policies for high-throughput proteomics, for example, by providing science-based evidence on which kinds of randomised decoy searches should be applied for reporting false identification rates in journal papers (Bradshaw *et al.*, 2006, Tabb, 2008). Future work to examine the decoy database performance on different datasets and pipelines will help to confirm the recommendations made from this study.

Chapter 4 also provided a tangible example of how novel research may be performed by exploiting datasets made public in a proteomics repository on the internet, such as those reviewed in Chapter 3. Mitchell Waldrop, a journalist at Scientific American, posed the question in April 2008 "*...is posting results online for all to see a great tool or a great risk?*"¹⁰⁷. The author argues that the work presented shows

¹⁰⁷ From an article entitled 'Science 2.0 Is open access science the future?'

that it is indeed worthwhile to make data freely available to the community via public proteomics repositories, such as Tranche at ProteomeCommons (as was used in Chapter 4). In this case there are more benefits for making datasets available than risks, especially since datasets are, in any case, usually only made available after publication. At present, one of the issues for proteomics data-sharing in this way is that there is currently more value placed on receiving citations for *papers* than there is value in having one's datasets re-applied and cited. This situation is likely to be improved with projects such as MIBBI (Minimum Information for Biological and Biomedical Investigations) (Taylor *et al.*, 2008). As Chis Taylor - who runs the project - stresses, external referencing of datasets that are obtained via repositories is important to incentivise authors to submit their data to databases in the first place (personal communication, July 2009). A formal system to make this possible is much needed. Moreover, by making data available in repositories, it is good for science as a whole, because it means reviewers can determine if the data is as good as authors say before work is accepted for publication. It also allows data to be re-used in new analyses, such as with new algorithms as they become available, and can lead to new tool development that is only possible when a critical mass of data is put together, such as consensus spectral searching, for example.

In Part III a new tool (MRMaid) and database (MRMaid-DB) were developed to support MRM studies. These systems, unlike those in Part II, focused on quantitative proteomics, an area of proteomics which is increasingly important permitting

detection of changes in the absolute quantities of proteins for use in systems biology and biomarker discovery and validation. Indeed, MRM is an effective technique for this purpose, but its limitation is that it is exclusively for *targeted* protein quantification; it must be hypothesis-driven. Other quantitative methods, such as label-free and those reviewed in Chapter 1 are required in situations where the proteins to target are not known - where there is no hypothesis to direct the quantification.

Designing transitions for SRM assays from scratch, or locating existing transitions in research papers is time-consuming, and does not leverage existing knowledge and data resources. MRMAid and MRMAid-DB offered new bioinformatics solutions to solve these problems. Indeed, since releasing MRMAid, several new publicly available tools for transition design and optimisation have subsequently emerged.

These new computational resources fall into two main types:

- (1) Web-based data-mining applications that sit between the user and a large public proteomics repository on the internet like MRMAid, they include The Global Proteome Machine's MRM Worksheet (Walsh *et al.*, 2009), and ESPPredictor (Fusaro *et al.*, 2009)
- (2) Standalone packages to predict transition candidates, which are installed and executed locally, such as MRMer (Martin *et al.*, 2008), Skyline (Prakash *et al.*, 2009), and MaRiMba (Sherwood *et al.*, 2009). This second group can exploit data from public repositories for the prediction process, for example via a web service or by applying specific spectral libraries.

The array of new offerings demonstrates how quickly the MRM approach is gaining popularity. Moreover, they highlight how appropriate the development of MRMAid was at the time. Indeed, MRMAid (along with TIQAM) lead the field of MRM informatics, being the first freely available tools in the world for transition design. Indeed, MRMAid has already been cited by others in recent papers and reviews, for

example in the MaRiMba paper (Sherwood *et al.*, 2009), and in a review article on MRM for clinical applications (Kim and Kim, 2009). To further contribute to the state of the art in this field, the author has written a review of publicly available software for MRM informatics. A review of these resources was desirable, since like repositories, the field is fast moving and researchers find it difficult to weigh up the capabilities and benefits of each offering. The review has been accepted by Proteomics journal and will be published as part of the HUPO conference proceedings edition of Proteomics in 2010 (see Appendix VII).

Also, MRMAid-DB has implications for the new standard format, TraML, for reporting transition data in the community. The process that the author went through to obtain all of the data items required to reproduce transitions is the same process that the HUPO-PSI work group are going through now. Therefore, it may be argued that by comparing the data-fields captured by MRMAid-DB with TraML, the MRMAid-DB schema may serve to validate the new standard going forward.

In summary, the outlook for MRMAid and MRMAid-DB is promising. Indeed, a new proposal is being put together in Autumn 2009 to obtain funding¹⁰⁸ to, among other things, integrate the MRMAid algorithm and its transition scoring, as well as the MRMAid-DB schema, into PRIDE database at the EBI. PRIDE is the only major public proteomics repository still to provide transition design functionality. Applying MRMAid in this way would provide a way to achieve longevity of support and continued global reach for the resources developed on this EngD. In general, a

¹⁰⁸From the bioinformatics and biological resources (BBR) fund from the BBSRC (Biotechnology and Biological Sciences Research Council).

major issue for tools like MRMAid is that they never progress from the level of 'prototype' to a robust application, and down-time on small-scale development servers is inevitable. By incorporating the MRM tool and database into PRIDE, which has professional support and regular sources of funding, there is the best chance of delivering benefits to users on a continued basis. Moreover, as part of the EBI's toolbox, there is greater likelihood that the proposed future improvements to MRMAid will go ahead (such as those mentioned at the end of Chapter 5).

Finally, it is likely that once SOPs are in place for the MRM approach, such as those that will be delivered by CPTAC, MRM will form a major part of the proteomics toolbox in both research environments and clinics. This means that sources of reliable transition candidates, such as MRMAid and MRMAid-DB, will become increasingly important in the future.

7.1 Wider opportunities for future work

To conclude, some further areas of research work are suggested to build on the achievements of this thesis.

7.1.1 Decoy database design options

The decoy databases created in Chapter 4 were based on randomising protein sequence databases based on known mathematical techniques, such as shuffling or reversing. There are other approaches to this randomisation that could be applied; for example, in proteins there are recurring biological motifs, the sequences are not completely random strings. New approaches to design decoy databases can take the

information on biological sequence motifs into account for decoy database creation, such as in a study that applied pattern recognition approaches and a Monte Carlo sampling algorithm (Feng *et al.*, 2007). If the statistical properties of the target are to be accurately mirrored in the decoy, then short repeating domains should also be accounted for in the decoy design. This avenue should be explored, and may lead to improved FPRs.

7.1.2 Refactoring of the systems: GAPP and MRMAid

Future work should include re-coding of GAPP and MRMAid from scratch, because there are several aspects that need to be improved. The original release of GAPP (Shadforth, 2005) was not well-documented and written in Perl. Issues with it include speed, presence of bugs in the code, and that the input and output formats are not PSI standards-compliant. Perhaps more importantly, however, is that the GAPP framework is based on Ensembl gene identifiers, hence MRMAid had to be based on these. Ensembl genes are not useful for protein isoforms, since each gene has a single ID even if several isoform variants exist. For MRMAid in particular this is a problem, because ideally users need to be able to monitor specific variants by MRM.

To make all of these changes in a piecemeal fashion would be a difficult task, and it would be more time-efficient to implement complete refactoring of the system, with the focus on delivering modularity and extensibility in the new version. Indeed, this would allow GAPP to become a robust application, not a prototype as it effectively

is now. Furthermore, a stand-alone version could be created at the same time and commercialisation explored.

7.1.3 Automated data harvesting

As explained in this thesis, there is a growing body of MS data available in public repositories. It is suggested, therefore, that changes be made to GAPP to allow harvesting of data from two major proteomic repositories, PRIDE and Tranche. One of the major pieces of work for the author in Chapter 5 was making sure enough spectral data was in GAPP DB, so MRMAid could make meaningful transition predictions. To improve MRMAid, therefore, more data needs to be made available to it via GAPP DB. To implement data-harvesting is not a simple operation because the minimum set of metadata will be required as search parameters for X!Tandem; hence, a middleware interface is required to connect to the repositories every time new data becomes available. The multiplicity of file formats for the spectra is also an issue to be accounted for, although it is hoped that HUPO-PSI's mzML will become the widely-adopted format of choice for MS/MS datasets in the near future.

The danger of taking data from multiple distributed sources is that some of it may be erroneous, or of poor quality. It is therefore important that some form of quality assessment be made on data presented to GAPP via automated harvesting. Various techniques to assess and improve spectral quality have already been described, for example in (Flikka *et al.*, 2006). Filters could be used to eliminate poor quality data prior to submission. In addition, post - processing could be developed to remove

errant identifications; for example, a rule-based expert system could apply knowledge of MS fragmentation chemistry taken from experts to identify and remove erroneous identifications, such as through knowledge of fragmentation chemistries and the limits of detection of MS.

7.1.4 Facility to execute different search engines in GAPP

To improve the data and identifications available to MRMAid further, and to make the score for optimising decoy database designs wider, it is suggested that GAPP pipeline be extended to include peptide scoring from others search engines in addition to X!Tandem including, for example, OMSSA and Mascot (for which the user would need to link GAPP to their Mascot server). This will also allow consensus scoring, whereby results from multiple search engines are combined to increase confidence, thus only identifications validated by searching various approaches would be stored in GAPP DB. It would also provide scope to combine the results of the different search engines into a single score, such as FDRScore as described in (Jones *et al.*, 2008a). Of course, Mascot is a commercially licensed software product so cannot be integrated into the open source distribution of GAPP, or made freely available as a part of the web-based installation of GAPP. However, users could 'farm out' processing to their own in-house Mascot server, if they have one. Mascot is a widely used and accepted search engine, particularly in the UK, so it would be a valuable addition for the purposes of testing decoy designs and for MRMAid.

7.2 Final word: proteomics gear-heads are here to stay

Biologists may feel frustrated by bioinformatics, believing that 'black boxes' (tools) developed by 'gear-heads' (bioinformaticians) do not represent 'real biology' research; rather a complicated means to an end. A recent sociology workshop explored this point; McNally - who lead it - found that biologists saw the growth in proteomic bioinformatics (the 'gear-head moment') merely as a transient stage in the history of biology research, and they *"look forward to the day when the gear-heads move aside to make room for biology to return to its rightful position centre stage and resume the 'real' job of science, which is addressing biological questions and meeting urgent social needs"* (McNally, 2008).

McNally's findings do not, however, reflect the reception the author has received for the work presented in this thesis. In fact, the author would argue that each piece of research presented here was designed specifically to *assist* lab-based researcher's endeavours by providing a means for researchers to leverage expertise and data from others, and to increase access and recognition to their own and others' work. Given the explosion in data volume, it is also hard to agree that proteomic bioinformatics research represents a short transient stage.

In reality, it is likely that even more sophisticated 'black boxes' will be needed to federate disparate data for the systems biology models of the future. It is more likely that the 'real biology' of the future will have computing as its central, novel focus, no

longer with the focus on the actual process of biological data capture, but rather the analysis afterwards. Evidence of this shift is already here: the CEO of the BBSRC, for example, concedes that increasing amounts of text and data are likely to change the entire epistemology of much of science¹⁰⁹. Indeed, it appears that the proteomics gear-heads are here to stay and the future of proteomics, and biology in general, needs them.

¹⁰⁹ Taken from Professor Douglas Kell's blog at the BBSRC website (<http://blogs.bbsrc.ac.uk/>), 'Computational infrastructure for modern biology' posted on 21st September, 2009.

References

- ALVES, G., WU, W. W., WANG, G., SHEN, R.-F. & YU, Y.-K. (2008) Enhancing Peptide Identification Confidence by Combining Search Methods. *J Proteome Res*, 7, 3102-3113.
- AMATORI, F. & JONES, G. (2003) *Business History Around the World*, Cambridge, Cambridge University Press.
- ANDERSON, L. & HUNTER, C. L. (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol Cell Proteomics*, 5, 573-588.
- ANDERSON, N. L., ANDERSON, N. G., HAINES, L. R., HARDIE, D. B., OLAFSON, R. W. & PEARSON, T. W. (2004) Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *J Proteome Res.*, 3, 235-44.
- ANDERSON, N. L., JACKSON, A., SMITH, D., HARDIE, D., BORCHERS, C. & PEARSON, T. W. (2009) SISCAPA peptide enrichment on magnetic beads using an inline beadtrap device. *Mol Cell Proteomics*, Feb 4. [Epub ahead of print]Click here to read M800446-MCP200.
- ANDERTON, S. M. (2004) Post-translational modifications of self antigens: implications for autoimmunity *Current Opinion in Immunology*, 16, 753-758.
- ANDREWS, P. C., ARNOTT, D. P., GAWINOWICZ, M. A., KOWALAK, J. A., LANE, W. S., LILLEY, K. S., MARTIN, L. T. & STEIN, S. E. (2006) ABRF-sPRG2006 study: a proteomics standard. *Poster available at <http://www.abrf.org/ResearchGroups/ProteomicsStandardsResearchGroup/EPosters/ABRFsPRGStudy2006poster.pdf>*, E-published.
- ANNAN, R. (2002) GlaxoSmithKline s Roland Annan Discusses the Phosphoproteome. *Proteomonitor*.
- ARLINGTON, S. (2007) What will pharma look like in 2020? *Drug Discovery World*, Summer, 9-16.
- ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S.,

- MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. (2000) Gene Ontology: tool for the unification of biology. *Nat Genet*, 25, 25-29.
- BALGLEY, B. M., LAUDEMAN, T., YANG, L., SONG, T. & LEE, C. S. (2007) Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. *Mol Cell Proteomics*, 6, 1599-608.
- BARINAGA, M. (1989) Protein chemists gain a new analytical tool. *Science*, 6, 32-33.
- BARNIDGE, D. R., DRATZ, E. A., MARTIN, T., BONILLA, L. E., MORAN, L. B. & LINDALL, A. (2003) Absolute Quantification of the G Protein-Coupled Receptor Rhodopsin by LC/MS/MS Using Proteolysis Product Peptides and Synthetic Peptide Standards. *Anal Chem*, 75, 445 -451.
- BARR, J. R., MAGGIO, V. L., PATTERSON, D. G., JR., COOPER, G. R., HENDERSON, L. O., TURNER, W. E., SMITH, S. J., HANNON, W. H., NEEDHAM, L. L. & SAMPSON, E. J. (1996) Isotope dilution--mass spectrometric quantification of specific proteins: model application with apolipoprotein A-I. *Clinical Chem*, 42, 1676-1682.
- BARSNES, H., VIZCAINO, J. A., EIDHAMMER, I. & MARTENS, L. (2009) PRIDE Converter: making proteomics data-sharing easy. *Nat Biotech*, 27, 598-599.
- BARTON, C., BECK, P., KAY, R., TEALE, P. & ROBERTS, J. (2009) Multiplexed LC-MS/MS analysis of horse plasma proteins to study doping in sport. *Proteomics*, 9, 3058-3065.
- BERRY, S. (2002) Biotech meets the investors. *Trends in Biotech*, 20, 370-371.
- BEYNON, R. J., DOHERTY, M. K., PRATT, J. M. & GASKELL, S. J. (2005) Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat Meth*, 2, 587-589.
- BLACKLER, A. R., KLAMMER, A. A., MACCOSS, M. J. & WU, C. C. (2006) Quantitative Comparison of Proteomic Data Quality between a 2D and 3D Quadrupole Ion Trap. *Anal Chem*, 78, 1337-1344.

- BLOW, N. (2008) Mass spectrometry and proteomics: hitting the mark. *Nat Meth*, 5, 741-747.
- BRADSHAW, R. A., BURLINGAME, A. L., CARR, S. & AEBERSOLD, R. (2006) Editorial: Reporting Protein Identification Data, The next Generation of Guidelines *MCP*, 5, 787-88.
- BRUNETTI, S., LODI, E., MORI, E. & STELLA, M. (2008) PARPST: a PARAllel algorithm to find peptide sequence tags. *BMC Bioinformatics*, 9 (Suppl 4), S11.
- BRUSH, C. G., GREENE, P. G. & HART, M. M. (2001) From initial idea to unique advantage: the entrepreneurial challenge of constructing a resource base. *Academy of Management Executive*, 15, 64-78.
- BUTLER, D. (2000) Celera in talks to launch private sector human proteome project. *Nature*, 403, 815-816.
- CARGILE, B. J., BUNDY, J. L. & STEPHENSON, J. L. J. (2004) Potential for false positive identifications from large databases through tandem mass spectrometry. *J Proteome Res.*, 3, 1082-5.
- CHEPANOSKE, C. L., RICHARDSON, B. E., VON RECHENBERG, M. & PELTIER, J. M. (2005) Average peptide score: a useful parameter for identification of proteins derived from database searches of liquid chromatography/tandem mass spectrometry data. *Rapid Comm Mass Spectrom*, 19, 9-14.
- CHOI, H. & NESVIZHISKII, A. I. (2008a) False Discovery Rates and Related Statistical Concepts in Mass Spectrometry-Based Proteomics. *J Proteome Res*, 7, 47-50.
- CHOI, H. & NESVIZHISKII, A. I. (2008b) Semisupervised Model-Based Validation of Peptide Identifications in Mass Spectrometry-Based Proteomics. *J Proteome Res*, 7, 254-265.
- CICCIMARO, E., HEVKO, J. & BLAIR, I. A. (2006) Analysis of phosphorylation sites on focal adhesion kinase using nanospray liquid chromatography/multiple reaction monitoring mass spectrometry. *Rapid Comm Mass Spectrom*, 20, 3681-3692.

- CLAMP, M., FRY, B., KAMAL, M., XIE, X., CUFF, J., LIN, M. F., KELLIS, M., LINDBLAD-TOH, K. & LANDER, E. S. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *PNAS USA*, 104, 19428-33.
- COLINGE, J., MASSELOT, A., GIRON, M., DESSINGY, T. & MAGNIN, J. (2003) OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics*, 3, 1454-1463.
- CÔTÉ, R. G., JONES, P., APWEILER, R. & HERMJAKOB, H. (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 28, 97.
- CÔTÉ, R. G., JONES, P., MARTENS, L., KERRIEN, S., REISINGER, F., LIN, Q., LEINONEN, R., APWEILER, R. & HERMJAKOB, H. (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*, 18, 401.
- COTTINGHAM, K. (2008) The PROSPECTS of analyzing proteomes in space and time. *Journal of Proteome Research*, 8, 429-429.
- COTTRELL, J. (2003) Matrix Science Founder John Cottrell Discusses the Business of Mascot. *Proteomonitor*.
- COX, D. M., ZHONG, F., DU, M., DUCHOSLAV, E., SAKUMA, T. & MCDERMOTT, J. C. (2005) Multiple Reaction Monitoring as a Method for Identifying Protein Posttranslational Modifications. *J. BioMol Techniques*, 16, 83-90.
- COX, J. & MANN, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotech*, 26, 1367-1372.
- CRAIG, R. & BEAVIS, R., C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass Spectrometry*, 17, 2310-2316.
- CRAIG, R., CORTENS, J. C., FENYO, D. & BEAVISA, R. C. (2006) Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. *J Proteome Res*, 5, 1843-1849.

- CRAIG, R., CORTENS, J. P. & BEAVIS, R. C. (2004) Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res*, 3, 1234-42.
- CRAIG, R., CORTENS, J. P. & BEAVIS, R. C. (2005) The use of proteotypic peptide libraries for protein identification. *Rapid Commun Mass Spectrom*, 19, 1844-50.
- DESIERE, F., DEUTSCH, E. W., KING, N. L., NESVIZHSHKII, A. I., MALLICK, P., ENG, J., CHEN, S., EDDES, J., LOEVENICH, S. N. & AEBERSOLD, R. (2006) The PeptideAtlas Project. *NAR*, 34, D655-D658.
- DEUTSCH, E. W., LAM, H. & AEBERSOLD, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows *EMBO Rep*, 9, 429-34.
- DIMASI, J. A., GRABOWSKI, H. G. & VERNON, J. (1995) R&D costs, innovative output and firm size in the pharmaceutical industry. *Int. Journal Econ Business*, 2, 201-219.
- DOMON, B. & AEBERSOLD, R. (2006) Mass Spectrometry and Protein Analysis. *Science*, 312, 212-217.
- DOWELL, R. D., JOKERST, R. M., DAY, A., EDDY, S. R. & STEIN, L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, 2, e-publication.
- DUNN, W. B. & ELLIS, D. I. (2005) Metabolomics: Current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry*, 24, 285-294.
- DWEK, R. A. (2008) Oxford University's first spin-off company - Oxford Glycosystems *The Biochemist*, April 2008, 4-7.
- EDITORIAL (1996) News: OXFORD GLYCOSCIENCES PLC *Analytical Instrument Industry Report*, 13.
- EDITORIAL (1997) ProteoGraph Oxford Glycosciences develops protein analysis technology. *R & D Focus Drug News*.
- EDITORIAL (1998) Oxford float - Oxford GlycoSciences. *The Times*. London.

EDITORIAL (1999) The promise of proteomics. *Nature*, 402, 703.

EDITORIAL (2007) Democratizing proteomics data. *Nat Biotech*, 25, 262.

EDITORIAL (2008a) The big ome. *Nature*, 452, 913-914.

EDITORIAL (2008b) Thou shalt share your data. *Nat Meth*, 5, 209.

EDMISTON, J., FLORA, J., SCIAN, M., LI, G., RANA, G., LANGSTON, T., SENGUPTA, T. & MCKINNEY, W. (2009) Cigarette smoke extract induced protein phosphorylation changes in human microvascular endothelial cells in vitro. *Analytical and Bioanalytical Chemistry*, 394, 1609-1620.

ELIAS, J. E. & GYGI, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Meth*, 4, 207-14.

EMBL (2007-08) EMBL Annual Report 2007-2008. Heidelberg, Germany, EMBL.

ENG, J. K., MCCORMACK, A. L. & YATES, J. R. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, 5, 976-989.

ERNST, W., TRUMMER, E., MEAD, J., BESSANT, C., STRELEC, H., KATINGER, H. & HESSE, F. (2006) Evaluation of a genomics platform for cross-species transcriptome analysis of recombinant CHO cells. *Biotechnology Journal*, 1, 639-650.

FALKNER, J. A., KACHMAN, M., VEINE, D. M., WALKER, A., STRAHLER, J. R. & ANDREWS, P. C. (2007) Validated MALDI-TOF/TOF Mass Spectra for Protein Standards. *J Am Soc Mass Spectrom*, 18, 850-5.

FALTH, M., SAVITSKI, M. M., NIELSEN, M. L., KJELDSEN, F., ANDREN, P. E. & ZUBAREV, R. A. (2007) SwedCAD, a Database of Annotated High-Mass Accuracy MS/MS Spectra of Tryptic Peptides. *J Proteome Res*, 6, 4063-4067.

FENG, J., NAIMAN, D. Q. & COOPER, B. (2007) Probability-based pattern recognition and statistical framework for randomization: modeling tandem

- mass spectrum/peptide sequence false match frequencies. *Bioinformatics*, 23, 2210-2217.
- FENN, J. B., MANN, M., MENG, C. K., WONG, S. F. & WHITEHOUSE, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246, 64-71.
- FIRN, D. (2000) Oxford Glyco shares up sharply on protein patents. *Financial Times*. 23rd March ed. London.
- FLIKKA, K., MARTENS, L., VANDEKERCKHOVE, J., GEVAERT, K. & EIDHAMMER, I. (2006) Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics*, 6, 2086-2094.
- FONG, T. (2009) At JPMorgan, Proteomics Tools Firms Lay Out Plans for Challenging 2009. *genomeweb.com*.
- FREWEN, B. & MACCOSS, M. J. (2007) Using BiblioSpec for creating and searching tandem MS peptide libraries. *Curr Protoc Bioinformatics*, 13.
- FROST&SULLIVAN (2001) World Proteomics Market Revenues To Hit \$6 Bln *Biomedical Market Newsletter*, 11.
- FUSARO, V. A., MANI, D. R., MESIROV, J. P. & CARR, S. A. (2009) Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat Biotech*, 27, 190-198.
- GARDEN, P., ALM, R. & HAKKINEN, J. (2005) PROTEIOS: an open source proteomics initiative. *Bioinformatics*, 21, 2085-2087.
- GARNIER, J.-P. (2008) Rebuilding the R&D engine in big pharma. *Harvard Business Rev*, May, 68-76.
- GEER, L. Y., MARKEY, S. P., KOWALAK, J. A., WAGNER, L., XU, M., MAYNARD, D. M., YANG, X., SHI, W. & BRYANT, S. H. (2004) Open Mass Spectrometry Search Algorithm. *J. Proteome Res.*, 3, 958-964.

- GERBER, S. A., RUSH, J., STEMMAN, O., KIRSCHNER, M. W. & GYGI, S. P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *PNAS*, 100, 6940-6945.
- GHOSH, D. & POISSON, L. M. (2009) "Omics" data and levels of evidence for biomarker discovery. *Genomics*, 93, 13-16.
- GORTZAK-UZAN, L., IGNATCHENKO, A., EVANGELOU, A. I., AGOCHIYA, M., BROWN, K. A., ST.ONGE, P., KIREEVA, I., SCHMITT-ULMS, G., BROWN, T. J., MURPHY, J., ROSEN, B., SHAW, P., JURISICA, I. & KISLINGER, T. (2008) A Proteome Resource of Ovarian Cancer Ascites: Integrated Proteomic and Bioinformatic Analyses To Identify Putative Biomarkers. *J Proteome Res*, 7, 339-351.
- GU, L., JONES, A. D. & LAST, R. L. (2007) LC-MS/MS Assay for Protein Amino Acids and Metabolically Related Compounds for Large-Scale Screening of Metabolic Phenotypes. 79, 8067-8075.
- GYGI, S. P., RIST, B., GERBER, S. A., TURECEK, F., GELB, M. H. & AEBERSOLD, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotech*, 17, 994-999.
- HAAS, W., FAHERTY, B. K., GERBER, S. A., ELIAS, J. E., BEAUSOLEIL, S. A., BAKALARSKI, C. E., LI, X., VILLEN, J. & GYGI, S. P. (2006) Optimization and Use of Peptide Mass Measurement Accuracy in Shotgun Proteomics. *Mol. Cell. Proteomics*, 5, 1326-1337.
- HAMILTON, D. P. & REGALADO, A. (2001) In hot pursuit of the proteome. *The Wall Street Journal*. New York.
- HAN, D. K., ENG, J., ZHOU, H. & AEBERSOLD, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat Biotech*, 19, 946-951.
- HANASH, S. M., STRAHLER, J. R., NEEL, J. V., HAILAT, N., MELHEM, R., KEIM, D., ZHU, X. X., WAGNER, D., GAGE, D. A. & WATSON, J. T. (1991) Highly resolving two-dimensional gels for protein sequencing. *PNAS*, 88, 5709-13.

- HARRISON, A. G., YOUNG, A. B., BLEIHOLDER, C., SUHAI, S. & PAIZS, B. (2006) Scrambling of Sequence Information in Collision-Induced Dissociation of Peptides. *J Am Soc Mass Spectrom*, 128, 10364-10365.
- HARTLER, J., THALLINGER, G. G., STOCKER, G., STURN, A., BURKARD, T. R., KÖRNER, E., RADER, R., SCHMIDT, A., MECHTLER, K. & TRAJANOSKI, Z. (2007) MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data. *BMC Bioinformatics*, 13, 197.
- HENCKE, D., EVANS, R. & RADFORD, T. (1999) Blair and Clinton push to stop gene patents. *The Guardian*. 20th Sept ed. London.
- HENZEL, W. J., BILLECI, T. M., STULTS, J. T., WONG, S. C., GRIMLEY, C. & WATANABE, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *PNAS*, 90, 5011-5015.
- HERMJAKOB, H. & APWEILER, R. (2006) The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert Rev Proteomics*, 3, 1-3.
- HIGDON, R., HOGAN, J. M., BELLE, G. V. & KOLKER, E. (2005) Randomized Sequence Databases for Tandem Mass Spectrometry Peptide and Protein Identification. *Omics J Integrative Biology*, 9, 364-379.
- HILLENKAMP, F. & KARAS, M. (1990) Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization. *Methods Enzymol*, 193, 280-95.
- HILLENKAMP, F., KARAS, M., BEAVIS, R. C. & CHAIT, B. T. (1991) Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical Chemistry*, 63, 1193A-1203A.
- HOOGLAND, C., MOSTAGUIR, K., SANCHEZ, J.-C., HOCHSTRASSER, D. F. & APPEL, R. D. (2004) SWISS-2DPAGE, ten years later. *Proteomics*, 4, 2352-2356.
- HOPP, T. P. & WOODS, K. R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *PNAS*, 78, 3824-3828.

- HUMMEL, J., NIEMANN, M., WIENKOOP, S., SCHULZE, W., STEINHAUSER, D., SELBIG, J., WALTHER, D. & WECKWERTH, W. (2007) ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites. *BMC Bioinformatics*, 8, 216-224.
- JAMES, P., QUADRONI, M., CARAFOLI, E. & GONNET, G. (1993) Protein Identification by Mass Profile Fingerprinting. *Biochemical and Biophysical Research Communications*, 195, 58-64.
- JONES, A. R., SIEPEN, J. A., HUBBARD, S. J. & PATON, N. W. (2008a) Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics*, 9, 1220 - 1229.
- JONES, P., CÔTÉ, R. G., CHO, S. Y., KLIE, S., MARTENS, L., QUINN, A. F., THORNEYCROFT, D. & HERMJAKOB, H. (2008b) PRIDE: new developments and new datasets. *NAR*, 36, D878-83.
- JONES, P., CÔTÉ, R. G., MARTENS, L., QUINN, A. F., TAYLOR, C. F., DERACHE, W., HERMJAKOB, H. & APWEILER, R. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *NAR*, 1, D659-63.
- JUNGBLUT, P. & WITTMANN-LIEBOLDB, B. (1995) Protein analysis on a genomic scale *Journal of Biotechnology*, 41, 111-120.
- KAHN, P. (1995) From Genome to Proteome: Looking at a Cell's Proteins. *Science*, 270, 369-370.
- KÄLL, L., STOREY, J. D., MACCOSS, M. J. & NOBLE, W. S. (2008) Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *J Proteome Res*, 7, 29-34.
- KAPP, E., SCHÜTZ, F., CONNOLLY, L., CHAKEL, J., MEZA, J., MILLER, C., FENYO, D., ENG, J., ADKINS, J., OMENN, G. & SIMPSON, R. (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *Proteomics*, 5, 3475-3490.

- KAY, R. G., BARTON, C., VELLOSO, C. P., BROWN, P. A., BARTLETT, C., BLAZEVIK, A. J., GODFREY, R. J., GOLDSPIK, G., REES, R., BALL, G. R., COWAN, D. A., HARRIDGE, S. D., ROBERTS, J., TEALE, P. & CREASER, C. S. (2009) High-throughput ultra-high-performance liquid chromatography/tandem mass spectrometry quantitation of insulin-like growth factor-I and leucine-rich a-2-glycoprotein in serum as biomarkers of recombinant human growth hormone administration. *Rapid Comm Mass Spectrom*, 23, 1-10.
- KAY, R. G., GREGORY, B., GRACE, P. B. & PLEASANCE, S. (2007) The application of ultra-performance liquid chromatography/tandem mass spectrometry to the detection and quantitation of apolipoproteins in human serum. *Rapid Comm Mass Spectrom*, 21, 2585-2593.
- KELLER, A., ENG, J., ZHANG, N., LI, X.-J. & AEBERSOLD, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol*, 1, E1-E8.
- KELLER, A., NESVIZHSHKII, A., KOLKER, E. & AEBERSOLD, R. (2002a) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, 74, 5383-92.
- KELLER, A., PURVINE, S., NESVIZHSHKII, A. I., STOLYAR, S., GOODLETT, D. R. & KOLKER, E. (2002b) Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis. *Omics J Integrative Biology*, 6, 207-212.
- KERNS, E. H. & DI, L. (2002) Utility of Mass Spectrometry for Pharmaceutical Profiling Applications. *Current Drug Metabolism* 7, 457-466
- KESHISHIAN, H., ADDONA, T., BURGESS, M., KUHN, E. & CARR, S. A. (2007) Quantitative, Multiplexed Assays for Low Abundance Proteins in Plasma by Targeted Mass Spectrometry and Stable Isotope Dilution. *Mol Cell Proteomics*, 6, 2212-29.
- KIM, K. & KIM, Y. (2009) Preparing multiple reaction monitoring for quantitative clinical proteomics *Expert Rev Proteomics*, 6, 225-229.
- KLAMMER, A. A. & MACCOSS, M. J. (2006) Effects of Modified Digestion Schemes on the Identification of Proteins from Complex Mixtures. *J Proteome Res*, 5, 695-700.

- KLIMEK, J., EDDES, J. S., HOHMANN, L., JACKSON, J., PETERSON, A., LETARTE, S., GAFKEN, P. R., KATZ, J. E., MALLICK, P., LEE, H., SCHMIDT, A., OSSOLA, R., ENG, J. K., AEBERSOLD, R. & MARTIN, D. B. (2008) The Standard Protein Mix Database: A Diverse Data Set To Assist in the Production of Improved Peptide and Protein Identification Software Tools. *J Proteome Res*, 7, 96-103.
- KOHLBACHER, O., REINERT, K., GROPL, C., LANGE, E., PFEIFER, N., SCHULZ-TRIEGLAFF, O. & STURM, M. (2007) TOPP--the OpenMS proteomics pipeline. *Bioinformatics*, 23, e191-197.
- KOVARIK, P., GRIVET, C., BOURGOGNE, E. & HOPFGARTNER, G. (2007) Method development aspects for the quantitation of pharmaceutical compounds in human plasma with a matrix-assisted laser desorption/ionization source in the multiple reaction monitoring mode. 21, 911-919.
- KROGH, A., LARSSON, B., VON HEIJNE, G. & SONNHAMMER, E. L. L. (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*, 305, 567-580.
- KROKHIN, O. V., CRAIG, R., SPICER, V., ENS, W., STANDING, K. G., BEAVIS, R. C. & WILKINS, J. A. (2004) An Improved Model for Prediction of Retention Times of Tryptic Peptides in Ion Pair Reversed-phase HPLC: Its Application to Protein Peptide Mapping by Off-Line HPLC-MALDI MS. *Mol Cell Proteomics*, 3, 908-919.
- KUHN, E., WU, J., KARL, J., LIAO, H., ZOLG, W. & GUILD, B. (2004) Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and ¹³C-labeled peptide standards. *Proteomics*, 4, 1175-1186.
- LAM, H., DEUTSCH, E. W., EDDES, J. S., ENG, J. K., KING, N., STEIN, S. E. & AEBERSOLD, R. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, 7, 655-667.
- LANE, W. S., NESVIZHSHKII, A. I., SEARLE, B., TABB, D. L., KOWALAK, J. A. & SEYMOUR, S. L. (2007) Bioinformatic Evaluation of Datasets Derived from the ABRF sPRG Proteomics Standard. *Poster available at <http://www.abrf.org/ResearchGroups/ProteomicsInformaticsResearchGroup/Studies/sPRG-BIC2007poster.pdf>*, E-published.

- LANGE, V., MALMSTROM, J., DIDION, J., KING, N. L., JOHANSSON, B. P., SCHAEFER, J., RAMESEDER, J., WONG, C.-H., DEUTSCH, E. W., BRUSNIAK, M.-Y., BUEHLMANN, P., BJOERCK, L., DOMON, B. & AEBERSOLD, R. (2008) Targeted quantitative analysis of *Streptococcus pyogenes* virulence factors by multiple reaction monitoring. *Mol Cell Proteomics*, 7, 1489-1500.
- LASHKARI, D. A., DERISI, J. L., MCCUSKER, J. H., NAMATH, A. F., GENTILE, C., HWANG, S. Y., BROWN, P. O. & DAVIS, R. W. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *PNAS USA*, 94, 13057-13062.
- LAU, K. W., JONES, A. R., SWAINSTON, N., SIEPEN, J. A. & HUBBARD, S. J. (2007) Capture and analysis of quantitative proteomic data. *Proteomics*, 7, 2787-2799.
- LENZ, C., KÜHN-HÖLSKEN, E. & URLAUB, H. (2007) Detection of Protein-RNA Crosslinks by Nano LC-ESI-MS/MS Using Precursor Ion Scanning and Multiple Reaction Monitoring (MRM) Experiments. *J. Am Soc Mass Spectrom*, 18, 869-881.
- LESTER, P. J. & HUBBARD, S. J. (2002) Comparative bioinformatic analysis of complete proteomes and protein parameters for cross-species identification in proteomics *Proteomics*, 2, 1392-1405.
- LI, X. J., ZHANG, H., RANISH, J. A. & AEBERSOLD, R. (2003) Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal Chem*, 75, 6648-57.
- LIU, P., CHENG, H., ROBERTS, T. M. & ZHAO, J. J. (2009) Targeting the phosphoinositide 3-kinase pathway in cancer. *Nat Rev Drug Discov*, 8, 627-644.
- LU, P., VOGEL, C., WANG, R., YAO, X. & MARCOTTE, E. M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotech*, 25, 117-124.
- MA, B., ZHANG, K., HENDRIE, C., LIANG, C., LI, M., DOHERTY-KIRBY, A. & LAJOIE, G. (2003) PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17, 2337-2342.

- MALLICK, P., SCHIRLE, M., CHEN, S. S., FLORY, M. R., LEE, H., MARTIN, D., RANISH, J., RAUGHT, B., SCHMITT, R., WERNER, T., KUSTER, B. & AEBERSOLD, R. (2007) Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotech.*, 25, 125-131.
- MANN, M. (1999) Quantitative proteomics? *Nat Biotech*, 17, 954-55.
- MANN, M., HØJRUP, P. & ROEPSTORFF, P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological Mass Spectrom*, 22, 338-345.
- MARCH, R. E. (1996) An introduction to quadrupole ion trap mass spectrometry. *J Mass Spectrometry*, 32, 351-369.
- MARKHAM, S. K. (2002) Moving technologies from lab to market. *Research Technology Management*, 45, 31-42.
- MARTENS, L., MÜLLER, M., STEPHAN, C., HAMACHER, M., REIDEGELD, K., A., MEYER, H., E., BLÜGGEL, 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.
- MARTIN, D. B., HOLZMAN, T., MAY, D., PETERSON, A., EASTHAM, A., ENG, J. & MCINTOSH, M. (2008) MRMer: An interactive open-source and cross-platform system for data extraction and visualization of multiple reaction monitoring experiments. *Mol Cell Proteomics*, Epub 30th June, M700504-MCP200.
- MARZOLF, B., DEUTSCH, E., MOSS, P., CAMPBELL, D., JOHNSON, M. & GALITSKI, T. (2006) SBEAMS-Microarray: database software supporting genomic expression analyses for systems biology. *BMC Bioinformatics*, 7, 286.
- MATHIVANAN, S., AHMED, M., AHN, N. G., ALEXANDRE, H., AMANCHY, R., ANDREWS, P. C., BADER, J. S., BALGLEY, B. M., BANTSCHIEFF, M., BENNETT, K. L., BJÖRLING, E., BLAGOEV, B., ... & PANDEY, A. (2008) Human Proteinpedia enables sharing of human protein data. *Nat Biotech*, 26, 164-5.

- MCKAY, M. J., SHERMAN, J., LAVER, M. T., BAKER, M. S., CLARKE, S. J. & MOLLOY, M. P. (2007) The development of multiple reaction monitoring assays for liver-derived plasma proteins. *Proteomics Clinical Applications*, 1, 1570-1581.
- MCKELVEY, M. (2000) *Evolutionary Innovations: The Business of Biotechnology*.
- MCLAUGHLIN, T., SIEPEN, J. A., SELLEY, J., LYNCH, J. A., LAU, K. W., YIN, H., GASKELL, S. J. & HUBBARD, S. J. (2006) PepSeeker: a database of proteome peptide identifications for investigating fragmentation patterns. *Nucl. Acids Res.*, 34, D649-654.
- MCMILLAN, J. (1994) Reorganizing vertical supply relationships. *Trends in Business Organization*, https://faculty-gsb.stanford.edu/mcmillan/personal_page/documents/Reorganizing%20Vertical%20Supply%20Relationships%201.pdf.
- MCNALLY, R. (2008) Sociomics: CESAGen Multidisciplinary Workshop on the Transformation of Knowledge Production in the Biosciences, and its Consequences 23-24 July, 2007 Wellcome Trust Genome Campus, Hinxton, UK. *Proteomics*, 8, 222-224.
- MEAD, J. A., BIANCO, L., OTTONE, V., BARTON, C., KAY, R. G., LILLEY, K. S., BOND, N. J. & BESSANT, C. (2009) MRMAid: the web-based tool for designing multiple reaction monitoring (MRM) transitions. *Mol Cell Proteomics*, 8, 696-705.
- MISHRA, G. R., SURESH, M., KUMARAN, K., KANNABIRAN, N., SURESH, S., BALA, P., SHIVAKUMAR, K., ANURADHA, N., REDDY, R., RAGHAVAN, T. M., MENON, S., HANUMANTHU, G., GUPTA, M., UPENDRAN, S., GUPTA, S., MAHESH, M., JACOB, B., MATHEW, P., CHATTERJEE, P., ARUN, K. S., SHARMA, S., CHANDRIKA, K. N., DESHPANDE, N., PALVANKAR, K., RAGHAVNATH, R., KRISHNAKANTH, R., KARATHIA, H., REKHA, B., NAYAK, R., VISHNUPRIYA, G., KUMAR, H. G., NAGINI, M., KUMAR, G. S., JOSE, R., DEEPTHI, P., MOHAN, S. S., GANDHI, T. K., HARSHA, H. C., DESHPANDE, K. S., SARKER, M., PRASAD, T. S. & PANDEY, A. (2006) Human protein reference database--2006 update. *NAR*, 1, D411-4.
- MITCHELL, P. (2003) In the pursuit of industrial proteomics. *Nat Biotech*, 21, 233-7.

- MOODY, G. (2004) *Digital Code of Life*, Hoboken, NJ, John Wiley & Sons.
- MOORE, R. E., YOUNG, M. K. & LEE, T. D. (2002) Qscore: an algorithm for evaluating SEQUEST database search results. *J Am Soc Mass Spectrom*, 13, 378-386.
- MUNDY, A. (2009) Drug Makers Fight Stimulus Provision. *The Wall Street Journal*. 10th Feb 2009 ed. New York.
- MURRAY, K. K., BOYD, R. K., EBERLIN, M. N., LANGLEY, G. J., LI, L., NAITO, Y. & TABET, J. C. (2005) IUPAC Standard Definitions of Terms Relating to Mass Spectrometry. *IUPAC MS Terms and Definitions, First Public Draft*.
- NAYLOR, S., CULBERTSON, A. W. & VALENTINE, S. J. (2007) Technology bane or bonanza for the pharmaceutical industry? *Drug Discovery World*, Fall, 53-58.
- NESVIZHISKII, A., KELLER, A., KOLKER, E. & AEBERSOLD, R. (2003a) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, 75, 4646-58.
- NESVIZHISKII, A. I. & AEBERSOLD, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics*, 4, 1419-1440.
- NESVIZHISKII, A. I., KELLER, A., KOLKER, E. & AEBERSOLD, R. (2003b) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75, 4646-4658.
- NESVIZHISKII, A. I., ROOS, F. F., GROSSMANN, J., VOGELZANG, M., EDDER, J. S., GRUISSEM, W., BAGINSKY, S. & AEBERSOLD, R. (2006) Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data: Toward More Efficient Identification of Post-translational Modifications, Sequence Polymorphisms, and Novel Peptides. *Mol Cell Proteomics*, 5, 652-670.
- NESVIZHISKII, A. I., VITEK, O. & AEBERSOLD, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Meth*, 4, 787-97.
- NEWSLINK (2003) 'Smart' companies get their rewards *Newcastle University's Newslink*.

- OMENN, STATES, ADAMSKI, BLACKWELL, MENON, HERMJAKOB, APWEILER, HAAB, SIMPSON, EDDER, KAPP, MORITZ, CHAN, RAI, ADMON, AEBERSOLD, ENG, HANCOCK, HEFTA, MEYER & AL, E. (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.
- ONG, S.-E., BLAGOEV, B., KRATCHMAROVA, I., KRISTENSEN, D. B., STEEN, H., PANDEY, A. & MANN, M. (2002) Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol Cell Proteomics*, 1, 376-386.
- ONISKO, B., DYNIN, I., REQUENA, J. R., SILVA, C. J., ERICKSON, M. & CARTER, J. M. (2007) Mass Spectrometric Detection of Attomole Amounts of the Prion Protein by nanoLC/MS/MS. *Journal of the American Society for Mass Spectrometry*, 18, 1070-1079.
- PAPPIN, D. J. C., HOJRUP, P. & BLEASBY, A. J. (1993) Rapid Identification of Proteins by Peptide-Mass Fingerprinting. *Curr Biology*, 3, 327-332.
- PAUL, V. W. & STEINWEDEL, H. (1953) Ein neues Massenspektrometer ohne Magnetfeld. *Zeitschrift fur Naturforschung*, 8a, 448-450.
- PENG, J., ELIAS, J. E., THOREEN, C. C., LICKLIDER, L. J. & GYGI, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res.*, 2, 43-50.
- PERKINS, D., PAPPIN, D., CREASY, D. & COTTRELL, J. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20, 3551-3567.
- PETRITIS, K., KANGAS, L. J., FERGUSON, P. L., ANDERSON, G. A., PASA-TOLIC, L., LIPTON, M. S., AUBERRY, K. J., STRITTMATTER, E. F., SHEN, Y., ZHAO, R. & SMITH, R. D. (2003) Use of Artificial Neural Networks for the Accurate Prediction of Peptide Liquid Chromatography Elution Times in Proteome Analyses. 75, 1039-1048.

- PICOTTI, P., LAM, H., CAMPBELL, D., DEUTSCH, E. W., MIRZAEI, H., RANISH, J., DOMON, B. & AEBERSOLD, R. (2008) A database of mass spectrometric assays for the yeast proteome. *Nat Meth*, 5, 913-14.
- PILLING, D. (2000) Clinton statement revitalises biotech sector *Financial Times*. 6th April ed. London.
- POLLACK, A. (2001) 3 Companies Will Try to Identify All Human Proteins. *The New York Times*. NY.
- PRAKASH, A., TOMAZELA, D. M., FREWEN, B., MACLEAN, B., MERRIHEW, G., PETERMAN, S. & MACCOSS, M. J. (2009) Expediting the Development of Targeted SRM Assays: Using Data from Shotgun Proteomics to Automate Method Development. *J Proteome Res.*, 0.
- PRASAD, B. N. R. (2004) Big Pharma R&D: is it worth spending? *European Case Clearing House*.
- PRESSRELEASE (2000a) GeneBio Appointed as Exclusive Worldwide Distributor of Proteome Systems' Glycosylation Database *PR Newswire*.
- PRESSRELEASE (2000b) Geneva Proteomics Raises Over \$40 Million in Financing *PR Newswire*. Evanston and Geneva.
- PRESSRELEASE (2000c) Novartis and Geneva Proteomics Establish Strategic Proteomics Alliance *PR Newswire Europe*. Basel and Geneva.
- PRESSRELEASE (2001a) Cellzome Acquires GlaxoSmithKline's Cell Map Unit. *PRNewsWire*.
- PRESSRELEASE (2001b) Index Ventures holds first closing on \$300 million fund. Names Life - Science Partner. . *Hugin Press Release*
- PRESSRELEASE (2001c) Marconi and OGS Form Joint Venture to Provide Proteomics Data And Hosting Solutions *PR Newswire*. London and Oxford.
- PRESSRELEASE (2001d) Myriad, Hitachi, Oracle & Friedli Join Forces to Map the Entire Human Proteome. *BusinessWire*. Salt Lake City.

PRESSRELEASE (2001e) OGS and the Institute for Systems Biology Announce ICAT based Proteomics Platform Alliance. *PR Newswire*. Oxford.

PRESSRELEASE (2001f) World's biggest proteomics plant opens in Geneva. *Reuters*. Geneva.

PRESSRELEASE (2003a) Celltech closes proteomics unit. *DrugResearcher.com*.

PRESSRELEASE (2003b) Information Technology: Alliance to develop next generation proteomics image software *Proteomics Weekly*.

PRESSRELEASE (2004a) Geneva Bioinformatics SA. *Instrument Business Outlook*, 18.

PRESSRELEASE (2004b) Matrix Science Establishes Subsidiary in Japan to Serve Growing Proteomics Market. *PRNewswire*.

PRESSRELEASE (2005a) Agreement inked to distribute updated Phenyx platform in France *Pharma Business Week*, 24.

PRESSRELEASE (2005b) Bayer HealthCare Diagnostics and Oxford Genome Sciences Partner in Proteomics-based Evaluation of New Biomarkers for Breast Cancer. *Oxford GenomeSciences*.

PRESSRELEASE (2005c) Distribution Agreement; Company announces major distribution agreement for Phenyx in Japan *Biotech Business Week*.

PRESSRELEASE (2005d) Expansion plans for Europe's leading bioinformatics institute. *Wellcome Trust Website*.

PRESSRELEASE (2005e) Nonlinear Dynamics announces more details of its global partnership with PerkinElmer. *Nonlinear Dynamics Website*.

PRESSRELEASE (2005f) Oxford Genome Sciences closes financing and moves to state-of-the-art proteomics facility. *Northbank Communications*.

PRESSRELEASE (2005g) Professional Education; Wiley-VCH to distribute proteomics-oriented bioinformatics training portal *Biotech Business Week*.

PRESSRELEASE (2005h) Sage-N Sorcerer(TM) Integrates ISB's Open Source Proteomics Tools; Plug-and-Play Appliance Provides Complete Analysis Workflow for High-Throughput Proteomics *Business Wire*.

PRESSRELEASE (2006a) Biosite and Oxford Genome Sciences Announce Collaboration in Colorectal Cancer. *PR Newswire US*.

PRESSRELEASE (2006b) Geneva Bioinformatics SA linked its Phenyx software with Insilicos' Proteomics Pipeline. *Instrument Business Outlook*.

PRESSRELEASE (2006c) GenoLogics partnered with Geneva Bioinformatics SA. *Instrument Business Outlook*.

PRESSRELEASE (2006d) Rosetta Biosoftware and Sage-N Research, Inc. Collaborate on Informatics Environment for Proteomics and Biomarker Discovery *Business Wire*.

PRESSRELEASE (2006e) Sage-N Research and GeneBio Sign North American Reseller Agreement *Business Wire*.

PRESSRELEASE (2006f) Scaffold Proteomics Data Mining Software Sold with Sage-N Sorcerer(TM) Systems for Low False Positive Protein ID Using Sequest(R) and Mascot(R) *Business Wire*.

PRESSRELEASE (2007a) Amgen and OGeS sign cancer Ab development deal *Pharma Marketletter*.

PRESSRELEASE (2007b) The Basel Biozentrum Taps Phenyx as Protein Identification Platform *Biotech Business Week*, 151.

PRESSRELEASE (2007c) Geneva Bioinformatics SA integrated its Phenyx MS data analysis software. *Instrument Business Outlook*.

PRESSRELEASE (2007d) Geneva Bioinformatics SA integrated Phenyx MS data analysis software with the Institute for Systems Biology's Trans Proteomic Pipeline. *Instrument Business Outlook*.

PRESSRELEASE (2007e) Nonlinear Dynamics expands distribution operation in the Asia Pacific region. *Nonlinear Dynamics Website*.

PRESSRELEASE (2007f) Oxford Genome Sciences (UK) Ltd secured additional private equity investment round. *Citigate Dewe Rogerson*.

PRESSRELEASE (2007g) Oxford Genome Sciences Announces New Therapeutic Antibody Deal With Medarex. *PR Newswire*.

PRESSRELEASE (2007h) Sage-N Research Gains Rights to Sell SEQUEST Proteomic Search Engine *Business Wire*.

PRESSRELEASE (2007i) Thirty years on: one human genome a day: Sanger Institute sequence production to increase 60-fold. *Wellcome Trust Sanger Institute*.

PRESSRELEASE (2008a) Biotech Business; Current BioData Opens Welsh Research and Publishing Center To Serve Global Pharmaceutical Industry *Biotech Business Week*, 3393.

PRESSRELEASE (2008b) GeneBio and Proxeon sign agreement for Phenyx and ProteinCenter. . *Company Reports*.

PRESSRELEASE (2008c) GlaxoSmithKline opens several fronts in campaign to adapt to a changing market *Pharma Marketletter*. London.

PRESSRELEASE (2008d) NCI Issues First Report on Proteomics Initiative; CPTAC Tackling Variability. *Proteomonitor*.

PRESSRELEASE (2008e) Nonlinear Dynamics to Distribute GeneBio's Phenyx Worldwide. *Nonlinear Website*.

PRESSRELEASE (2008f) Protagen and GeneBio sign distribution agreement for Modiro M2 *Europharma*.

PRESSRELEASE (2008g) Thermo Fisher Scientific and Sage-N Research introduced the SORCERER Enterprise software product. *Instrument Business Outlook*.

PRESSRELEASE (2009) GeneBio and Utrecht University Announce Collaborative Partnership Around Phenyx. *GeneBio website*.

QIAN, W. J., LIU, T., MONROE, M. E., STRITTMATTER, E. F., JACOBS, J. M., KANGAS, L. J., PETRITIS, K., CAMPII, D. G. & SMITH, R. D. (2005) Probability-Based Evaluation of Peptide and Protein Identifications from Tandem Mass Spectrometry and SEQUEST Analysis: The Human Proteome. *J. Proteome Res.*, 4, 53-62.

RAUCH, A., BELLEW, M., ENG, J., FITZGIBBON, M., HOLZMAN, T., HUSSEY, P., IGRA, M., MACLEAN, B., LIN, C. W., DETTER, A., FANG, R., FACA, V., GAFKEN, P., ZHANG, H., WHITAKER, J., STATES, D., HANASH, S., PAULOVICH, A. & MCINTOSH, M. W. (2006) Computational Proteomics Analysis System (CPAS): An Extensible, Open-Source Analytic System for Evaluating and Publishing Proteomic Data and High Throughput Biological Experiments. *J Proteome Res*, 5, 112-121.

REIDEGELD, K. A., EISENACHER, M., KOHL, M., CHAMRAD, D., KÖRTING, G., BLÜGGEL, M., MEYER, H. E. & STEPHAN, C. (2008) An easy-to-use Decoy Database Builder software tool, implementing different decoy strategies for false discovery rate calculation in automated MS/MS protein identifications. *Proteomics*, 8, 1129-1137.

REIDEGELD, K. A., HAMACHER, M., MEYER, H. E., STEPHAN, C., BLÜGGEL, M., KÖRTING, G., CHAMRAD, D., SCHEER, C., THIELE, H., TAYLOR, C., MÜLLER, M., APWEILER, R., JONES, P. & MARTENS, L. (2006) The HUPO brain proteome project. *European Pharmaceutical Review*, 33-38.

RIFAI, N., GILLETTE, M. A. & CARR, S. A. (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotech*, 24, 971-983.

ROEPSTORFF, P. & FOHLMANN, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry*, 11, 601.

ROHLFF, C. (2004) New approaches towards integrated proteomic databases and depositories. *Expert Rev Proteomics*, 1, 267-274.

RUSSELL, J. (2008) A Witty cure for a sick patient. *The Daily Telegraph*. London.

- SADYGOV, R., COCIORVA, D. & YATES, J., R. (2004) Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nat Meth*, 1, 195-202.
- SAHOO, A. (2006) Pharmaceutical outsourcing strategies. Market expansion, offshoring and strategic management in the CRO and CMO marketplace. *Business Insights*.
- SEARLE, B. C., DASARI, S., TURNER, M., REDDY, A. P., CHOI, D., WILMARTH, P. A., MCCORMACK, A. L., DAVID, L. L. & NAGALLA, S. R. (2004) High-Throughput Identification of Proteins and Unanticipated Sequence Modifications Using a Mass-Based Alignment Algorithm for MS/MS de Novo Sequencing Results. *Analytical Chemistry*, 76, 2220-2230.
- SEARLE, B. C., DASARI, S., WILMARTH, P. A., TURNER, M., REDDY, A. P., DAVID, L. L. & NAGALLA, S. R. (2005) Identification of Protein Modifications Using MS/MS de Novo Sequencing and the OpenSea Alignment Algorithm. *Journal of Proteome Research*, 4, 546-554.
- SEARLE, B. C., TURNER, M., Nesvizhskii, A. I. (2008) Improving Sensitivity by Probabilistically Combining Multiple MS/MS Search Methodologies. *Journal of Proteome Research*, 7, 245-253.
- SERVICE, R. F. (2000) Can Celera Do It Again? *Science*, 287, 2136 - 2138.
- SERVICE, R. F. (2008a) PROTEOMICS: Proteomics Ponders Prime Time. *Science*, 321, 1758-1761.
- SERVICE, R. F. (2008b) PROTEOMICS: Will Biomarkers Take Off at Last? *Science*, 321, 1760-.
- SHADFORTH, I. (2005) Development and implementation of improved peptide identification algorithms in a high-throughput proteomics pipeline *Department of Analytical Science and Informatics (DASI)*. Cranfield University.
- SHADFORTH, I. & BESSANT, C. (2006) Genome annotating proteomics pipelines: available tools. *Expert Rev. Proteomics*, 3.

- SHADFORTH, I., CROWTHER, D. & BESSANT, C. (2005a) Protein and peptide identification algorithms using MS for use in high-throughput, automated pipelines. *Proteomics*, 5, 4082-4095.
- SHADFORTH, I., DUNKLEY, T., LILLEY, K., CROWTHER, D. & BESSANT, C. (2005b) Confident protein identification using the average peptide score method coupled with search-specific, ab initio thresholds. *Rapid Communications in Mass Spectrometry*, 19, 3363-3368.
- SHADFORTH, I., XU, W., CROWTHER, D. & BESSANT, C. (2006) GAPP: A Fully Automated Software for the Confident Identification of Human Peptides from Tandem Mass Spectra. *J. Proteome Res.*, 5, 2849 -2852.
- SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*, 13, 2498-2504.
- SHERWOOD, C., EASTHAM, A., PETERSON, A., ENG, J. K., SHTEYNBERG, D., MENDOZA, L., DEUTSCH, E., RISLER, J., LEE, L. W., TASMAN, N., AEBERSOLD, R., LAM, H. & MARTIN, D. B. (2009) MaRiMba: a Software Application for Spectral Library-Based MRM Transition List Assembly. *Journal of Proteome Research*.
- SHEVCHENKO, A., JENSEN, O. N., PODTELEJNIKOV, A. V., SAGLIOCCO, F., WILM, M., VORM, O., MORTENSEN, P., SHEVCHENKO, A., BOUCHERIE, H. & MANN, M. (1996) Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels *PNAS*, 93, 14440-14445.
- SIEPEN, J. A., BELHAJJAME, K., SELLEY, J. N., EMBURY, S. M., PATON, N. W., GOBLE, C. A., OLIVER, S. G., STEVENS, R., ZAMBOULIS, L., MARTIN, N., POULOVASSILLIS, A., JONES, P., COTE, R., HERMJAKOB, H., PENTONY, M. M., JONES, D. T., ORENGO, C. A. & HUBBARD, S. J. (2008) ISPIDER Central: an integrated database web-server for proteomics. *NAR*, Epub, gkn196.
- SIEPEN, J. A., SWAINSTON, N., JONES, A. R., HART, S. R., HERMJAKOB, H., JONES, P. & HUBBARD, S. J. (2007) An informatic pipeline for the data capture and submission of quantitative proteomic data using iTRAQTM. *Proteome Sci*, 1, 4.

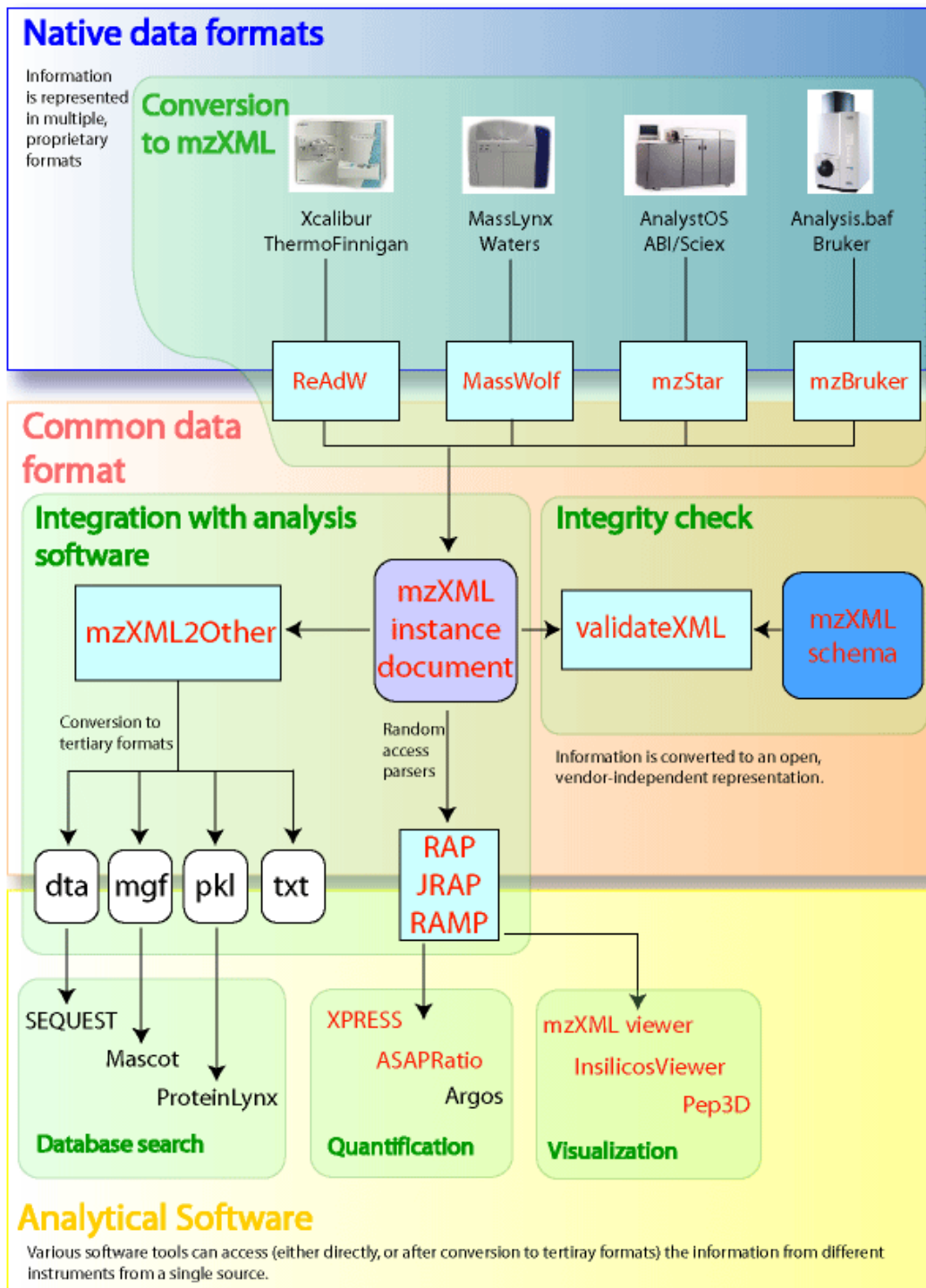
- SINGHAL, P., GAUR, A., GAUTAM, A., VARSHNEY, B., PALIWAL, J. & BATRA, V. (2007) Sensitive and rapid liquid chromatography/tandem mass spectrometric assay for the quantification of piperazine in human plasma. *Journal of Chromatography B*, 859, 24-29.
- STAHL-ZENG, J., LANGE, V., OSSOLA, R., ECKHARDT, K., KREK, W., AEBERSOLD, R. & DOMON, B. (2007) High Sensitivity Detection of Plasma Proteins by Multiple Reaction Monitoring of N-Glycosites. *Mol Cell Proteomics*, 6, 1809-1817.
- STEPHAN, C., REIDEGELD, K., A., HAMACHER, M., VAN HALL, A., MARCUS, K., TAYLOR, C., JONES, P., MÜLLER, M., APWEILER, R., MARTENS, L., KÖRTING, G., CHAMRAD, D., C., THIELE, H., BLÜGGEL, M., PARKINSON, D., BINZ, P.-A., LYALL, A. & MEYER, H., E. (2006) Automated reprocessing pipeline for searching heterogeneous mass spectrometric data of the HUPO Brain Proteome Project pilot phase. *Proteomics*, 6, 5015-5029.
- STURM, M., BERTSCH, A., GRÖPL, C., HILDEBRANDT, A., HUSSONG, R., LANGE, E., PFEIFER, N., SCHULZ-TRIEGLAFF, O., ZERCK, A., REINERT, K. & KOHLBACHER, O. (2008) OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics*, 9, 163-174.
- SUMMARY (2005) Biotechnology company deals with miscellaneous companies: collaborations, agreements, equity participation, June 17-Aug. 23, 2005. *BioWorld Financial Watch*.
- TABB, D. L. (2008) What's Driving False Discovery Rates? *J Proteome Res*, 7, 45-46.
- TABB, D. L., SARAF, A. & YATES, J. R. (2003) GutenTag: High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model. *Anal. Chem.*, 75, 6415-6421.
- TANAKA, K., WAKI, H., IDO, Y., AKITA, S., YOSHIDA, Y. & YOSHIDA, T. (1988) Protein and Polymer Analyses up to m/z 100 000 by Laser Ionization Time-of flight Mass Spectrometry. *Rapid Comm Mass Spectrom*, 2, 151-153.
- TANG, H., ARNOLD, R. J., ALVES, P., XUN, Z., CLEMMER, D. E., NOVOTNY, M. V., REILLY, J. P. & RADIVOJAC, P. (2006) A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics*, 22, e481-488.

- TAYLOR, C. F., FIELD, D., SANSONE, S.-A., AERTS, J., APWEILER, R., ASHBURNER, M., BALL, C. A., BINZ, P.-A., BOGUE, M., BOOTH, T., BRAZMA, A., BRINKMAN, R. R., MICHAEL CLARK, A., DEUTSCH, E. W., FIEHN, O., FOSTEL, J., GHAZAL, P., GIBSON, F., GRAY, T., GRIMES, G., HANCOCK, J. M., HARDY, N. W., HERMJAKOB, H., JULIAN, R. K., KANE, M., KETTNER, C., KINSINGER, C., KOLKER, E., KUIPER, M., NOVERE, N. L., LEEBENS-MACK, J., LEWIS, S. E., LORD, P., MALLON, A.-M., MARTHANDAN, N., MASUYA, H., MCNALLY, R., MEHRLE, A., MORRISON, N., ORCHARD, S., QUACKENBUSH, J., REECY, J. M., ROBERTSON, D. G., ROCCA-SERRA, P., RODRIGUEZ, H., ROSENFELDER, H., SANTOYO-LOPEZ, J., SCHEUERMANN, R. H., SCHOBER, D., SMITH, B., SNAPE, J., STOECKERT, C. J., TIPTON, K., STERK, P., UNTERGASSER, A., VANDESOMPELE, J. & WIEMANN, S. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotech*, 26, 889-896.
- TAYLOR, C. F., PATON, N. W., GARWOOD, K. L., KIRBY, P. D., STEAD, D. A., YIN, Z., DEUTSCH, E. W., SELWAY, L., WALKER, J., RIBA-GARCIA, I., MOHAMMED, S., DEERY, M. J., HOWARD, J. A., DUNKLEY, T., AEBERSOLD, R., KELL, D. B., LILLEY, K. S., ROEPSTORFF, P., YATES, J. R., BRASS, A., BROWN, A. J. P., CASH, P., GASKELL, S. J., HUBBARD, S. J. & OLIVER, S. G. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat Biotech*, 21, 247-254.
- TAYLOR, G. K. & GOODLETT, D. R. (2005) Rules governing protein identification by mass spectrometry. *Rapid Comm Mass Spectrom*, 19, 3420.
- THOMSEN, A. L., LAUKENS, K., MATTHIESEN, R. & JENSEN, O. N. (2007) Organization of proteomics data with YassDB. *Methods Mol Biol*, 367, 271-87.
- UETZ, P., GIOT, L., CAGNEY, G., MANSFIELD, T. A., JUDSON, R. S., KNIGHT, J. R., LOCKSHON, D., NARAYAN, V., SRINIVASAN, M., POCHART, P., QURESHI-EMILI, A., LI, Y., GODWIN, B., CONOVER, D., KALBFLEISCH, T., VIJAYADAMODAR, G., YANG, M., JOHNSTON, M., FIELDS, S. & ROTHBERG, J. M. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403, 623-627.
- UHLEN, M. & PONTEN, F. (2005) Antibody-based Proteomics for Human Tissue Profiling. *Mol Cell Proteomics*, 4, 384-393.

- UNWIN, R. D., GRIFFITHS, J. R., LEVERENTZ, M. K., GRALLERT, A., HAGAN, I. M. & WHETTON, A. D. (2005) Multiple reaction monitoring to identify sites of protein phosphorylation with high sensitivity. *Mol. Cell. Proteomics*, 4, 1134-1144.
- VAN DER SIJDE, P., BLIEK, P. & GROEN, A. (2003) The exploitation of biotech innovations: networking for survival and success. *European Case Clearing House*.
- WADE, N. (1981) The complete index to man. *Science*, 2, 33-35.
- WALSH, G. M., LIN, S., EVANS, D. M., KHOSROVI-EGHBAL, A., BEAVIS, R. C. & KAST, J. (2009) Implementation of a data repository-driven approach for targeted proteomics experiments by multiple reaction monitoring. *J Proteomics*, 72, 838-52.
- WASHBURN, M., WOLTERS, D. & YATES, J. R. R. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 19, 242-47.
- WASHTECH@WASHINGTONPOST (2000) Celera Genomics Tacks On More Stock For Parent. *The Washington Post*. Rockville, Maryland.
- WEBB-ROBERTSON, B.-J. M., CANNON, W. R., OEHMEN, C. S., SHAH, A. R., GURUMOORTHY, V., LIPTON, M. S. & WATERS, K. M. (2008) A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics*, 24, 1503-1509.
- WELLS, M. R., YOUNG, K. & FLEISHAUER, C. (2008) *Four Paws from Heaven: Inspirational Stories for Dog Lovers*, Eugene, Oregon, Harvest House Publishers.
- WOLF-YADLIN, A., HAUTANLEML, S., LAUFFENBURGER, D. A. & WHITE, F. M. (2007) Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *PNAS*, 104, 5860-5865.
- WU, Z., ASOKAN, A., GRIEGER, J. C., GOVINDASAMY, L., AGBANDJE-MCKENNA, M. & SAMULSKI, R. J. (2006) Single Amino Acid Changes Can Influence Titer, Heparin binding, and Tissue Tropism in Different Adeno-Associated Virus (AAV) Serotypes. *J Virol*, 80, 11393-7.

- XU, C. & MA, B. (2006) Software for computational peptide identification from MS-MS data. *Drug Discovery Today*, 11, 595-600.
- YATES, J. R., S, S., P.R., G. & T., H. (1993) Peptide mass maps: a highly informative approach to protein identification. *Anal Biochem*, 214, 397-408.
- YEN, C. Y., RUSSELL, S., MENDOZA, A. M., MEYER-ARENDR, K., SUN, S., CIOS, K. J., AHN, N. G. & RESING, K. A. (2006) Improving Sensitivity in Shotgun Proteomics Using a Peptide-Centric Database with Reduced Complexity: Protease Cleavage and SCX Elution Rules from Data Mining of MS/MS Spectra. *Anal Chem*, 78, 1071-1084.
- ZHANG, C., CRASTA, O., CAMMER, S., WILL, R., KENYON, R., SULLIVAN, D., YU, Q., SUN, W., JHA, R., LIU, D., XUE, T., ZHANG, Y., MOORE, M., MCGARVEY, P., HUANG, H., CHEN, Y., ZHANG, J., MAZUMDER, R., WU, C. & SOBRAL, B. (2008) An emerging cyberinfrastructure for biodefense pathogen and pathogen host data. *NAR*, 36, D884-891.
- ZHANG, F., BARTELS, M. J. & STOTT, W. T. (2004) Quantitation of human glutathione S-transferases in complex matrices by liquid chromatography/tandem mass spectrometry with signature peptides. *Rapid Comm Mass Spec*, 18, 491-498.
- ZHANG, G., FAN, H., XU, C., BAO, H. & YANG, P. (2003) On-line preconcentration of in-gel digest by ion-exchange chromatography for protein identification using high-performance liquid chromatography-electrospray ionization tandem mass spectrometry. *Analytical Biochemistry*, 313, 327-330.
- ZHANG, J., GAO, W., CAI, J., HE, S., ZENG, R. & CHEN, R. (2005) Predicting Molecular Formulas of Fragment Ions with Isotope Patterns in Tandem Mass Spectra. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2, 217-230.
- ZHANG, Y., ZHANG, Y., ADACHI, J., OLSEN, J. V., SHI, R., DE SOUZA, G., PASINI, E., FOSTER, L. J., MACEK, B., ZOUGMAN, A., KUMAR, C., WISNIEWSKI, J. R., JUN, W. & MANN, M. (2007) MAPU: Max-Planck Unified database of organellar, cellular, tissue and body fluid proteomes. *Nucl. Acids Res.*, 35, D771-779.

Appendix I



Appendix I Figure 1 A summary of software tools to interchange proteomics data formats (Source: ISB ProteomeCenter)

Appendix I Table 1 Short courses/ workshops attended during the EngD

Course	Course content
Research planning and report writing	Scientific report writing, data analysis and interpretation skills
Techniques to aid innovation	Design/realisation process, IP rights, statistical experiment design, new manufacturing processes
GRAD summer school	Personal development, teamwork, facilitation skills and confidence
Technology change and environmental assessment	Auditing methods for environmental impact
Learning team approach course	Team-building, leadership, inter-personal skills, problem-solving
Intelligent systems	Architecture, knowledge engineering and control, languages used in expert systems, Bayesian inference, fuzzy logic and decision support systems including clinical applications

Appendix I Table 2 Conferences, Presentations and Posters for the EngD

Date	Title of EngD work presented	Presentation type	Conference or meeting name	Location	Materials in appendix CD
21-25/07/07	Novel bioinformatics tools for cross-species data analysis	Poster	15 th International conference on Intelligent Systems for Molecular Biology, and 6 th European Conference on Computational Biology	Vienna, Austria	No
19/11/07	-	-	6 th Annual Proteomics Day and Second BSPR London Regional meeting	London, UK	-
6-7/05/08	Using Bioinformatics to Increase Speed and Reduce Uncertainty in Protein Biomarker Discovery	Oral and paper	Cranfield Multi-Strand Conference	Cranfield, UK	Paper
18-19/06/08	Bridging the GAPP	Oral	Proteomics Method Forum	Dundee, Scotland	No
8-10/07/08	Fast and reliable MRM transition design	Oral	British Society of Proteome Research (BSPR) / European Bioinformatics Institute (EBI) Conference [Proteomics: From technology to new biology]	Hinxton, UK	No
22-26/09/08	Applying community standard MS/MS datasets to evaluate proteomic data analysis pipeline performance	Poster	7 th European Conference on Computational Biology	Cagliari, Sardinia, Italy	Poster
01/12/08	-	-	The Seventh Annual Proteomics Day, and Third BSPR London Regional meeting	London, UK	-
20/05/08	Man versus MRmaid: can a computer program design transitions as well as a Quotient scientist?	Oral	Quotient BioResearch Ltd. lunchtime seminar	Fordham, Newmarket, UK	No
27/06/09-02/07/09	MRmaid: automating the design of multiple reaction monitoring (MRM) experiments using expert knowledge and MS/MS data-mining	Poster	17 th International conference on Intelligent Systems for Molecular Biology, and 8 th European Conference on Computational Biology	Stockholm, Sweden	Poster
14-16/07/09	Which decoy database gives the lowest false positive rate in automated searches using a proteomics pipeline?	Poster	British Society of Proteome Research (BSPR) / European Bioinformatics Institute (EBI) Conference [Multiscale proteomics: from cells to organisms]	Hinxton, UK	Poster

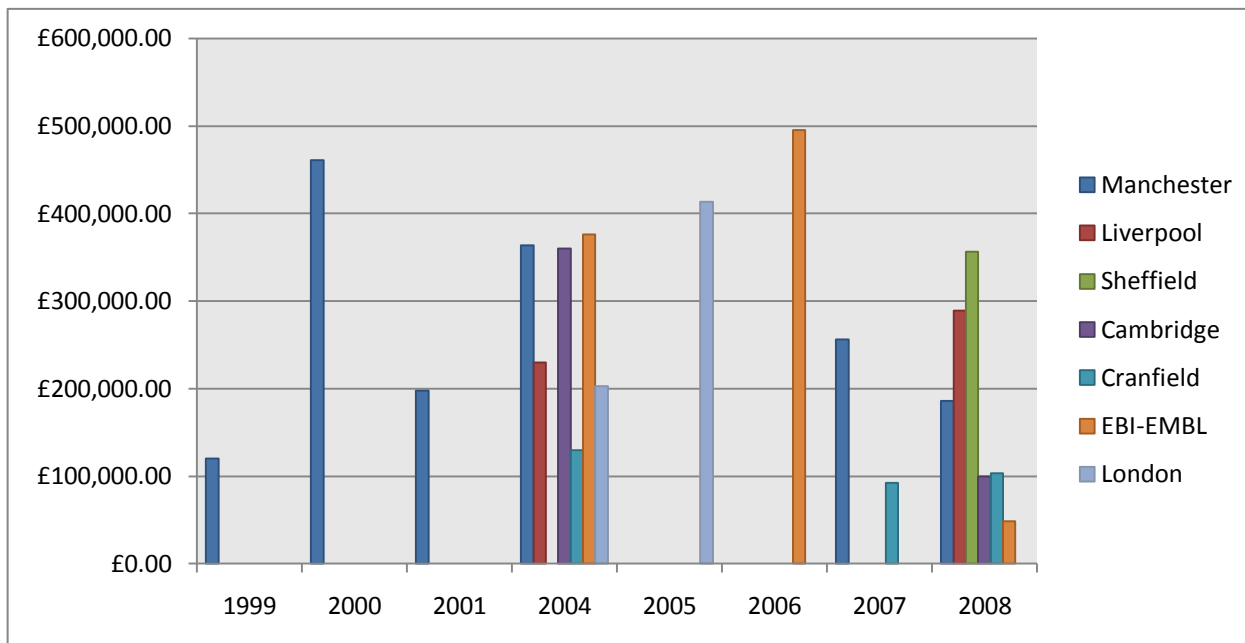
Appendix I Table 3 Professional memberships of the author, as a result of expertise gained on the EngD programme

Date joined	Organisation
2009	The American Chemical Society
2007	The International Society for Computational Biology
2007	The British Society for Proteome Research

Appendix II

Appendix II Table 1 BBSRC proteomic MS grants awarded between 1999 and 2008. The colours indicate linked grants, such as follow on grants. Asterisk indicates grants which also include additional proteomics approaches (in addition to MS)

Year	Start date	Grant	Proposal details	Recipient	Establishment
1999	11/01/1999	£119,898	Software tools for complex mixture analysis of proteome proteins	hubbard	manchester
2000	01/04/2000	£460,964	Making the most of a genome sequence: the application of global transcriptome and proteome analysis to <i>Streptomyces coelicolor</i> A3(2)*	smith	manchester
2001	01/08/2001	£197,912	Realising a qualitative increase in the capacity of proteomics by statistical image analysis of 2D electrophoresis gels	graham	manchester
2004	01/05/2004	£376,187	ISPIDER - a pilot grid for integrative proteomics	apweiler	EBI
2004	01/07/2004	£360,181	Developing PEDRo as a standard tool for the capture, representation, analysis, and dissemination of proteomics data	oliver	cambridge
2004	01/10/2004	£202,883	ISPIDER - a pilot grid for integrative proteomics	martin	birkbeck
2004	25/10/2004	£363,821	ISPIDER - a pilot grid for integrative proteomics	hubbard	manchester
2004	15/11/2004	£230,130	Application of field programmable gate arrays to eliminate bottlenecks in near-instrument proteomics	beynon	liverpool
2004	22/11/2004	£129,699	A Grid-Based System for Cataloguing the Human Proteome from Distributed Mass Spectrometry Data	bessant	cranfield
2005	01/04/2005	£222,889	ISPIDER - a pilot Grid for integrative proteomics	jones	UCL
2005	01/09/2005	£20,000	Development and Dissemination of e-Protein: A distributed pipeline for proteome annotation using GRID technology	jones	UCL
2005	01/09/2005	£19,965	Development and Dissemination of e-Protein - a Distributed Annotation Pipeline for Proteome annotation using Grid technology	sternberg	imperial
2005	01/12/2005	£150,193	Computing Equipment for Bioinformatics *	sternberg	imperial
2006	01/05/2006	£444,800	EMBOSS: European Molecular Biology Open Software Suite *	Rice	EMBL
2006	01/10/2006	£50,205	ProteomeHarvest - Excel/XML Bridge for User-friendly Proteomics Data Collection	Apweiler	EBI-EMBL
2007	01/04/2007	£92,274	Further Development of the Genome Annotating Proteomic Pipeline	bessant	cranfield
2007	01/05/2007	£156,781	A Multi-Processor Linux Farm for Bioinformatics and Functional Genomics *	lovell	manchester
2007	07/05/2007	£99,103	Informatics tools for analysis of quantitative proteomics data	hubbard	manchester
2008	14/01/2008	£186,111	Rapid proteome profiling using positional signature peptides	hubbard	manchester
2008	01/03/2008	£48,383	Database on demand - creating customized sequence databases for efficient protein identification	apweiler	EBI-EMBL
2008	14/03/2008	£71,545	FPGA supercomputing technology for high-throughput identification and quantitation in proteomics	beynon	liverpool
2008	15/03/2008	£356,480	FPGA supercomputing technology for high-throughput identification and quantitation in proteomics	coca	sheffield
2008	01/06/2008	£217,459	Rapid proteome profiling using positional signature peptides	beynon	liverpool
2008	01/07/2008	£103,094	X-tracker: a generic quantitation tool for MS-based proteomics	bessant	cranfield
2008	18/09/2008	£99,876	Computational methods to enable construction of 3D models of protein complexes by integrating mass spectrometry and biochemical data	robinson	cambridge
Total		£4,780,833			



Appendix II Figure 1 Breakdown of public funding awarded for proteomic bioinformatics research in England between 1999 and 2008.



7,276 visits came from 995 cities

Site Usage				
Visits 7,276 % of Site Total: 100.00%	Pages/Visit 4.79 Site Avg: 4.79 (0.00%)	Avg. Time on Site 00:05:02 Site Avg: 00:05:02 (0.00%)	% New Visits 44.80% Site Avg: 44.72% (0.18%)	Bounce Rate 60.34% Site Avg: 60.34% (0.00%)

Appendix II Figure 2 Hits for GPMDB as an indicator of activity in proteomics research across the world during June 2009

Appendix III

Paper 1 of the EngD

- Mead, J.A. and Shadforth, I.P. (2007) *Bringing protein identification to the masses. Institute of Biology Biologist 54:200-206*

paper1_2007_biologist.pdf

Paper 2 of the EngD

- Mead, J.A., Shadforth, I.P. and Bessant C. (2007) *Public proteomic MS repositories and pipelines: available tools and biological applications Proteomics 7(16): 2769-86*

paper2_2007_proteomics.pdf

Paper 3 of the EngD

- Mead, J.A., Bianco, L. and Bessant C. (2009) *Recent developments in public proteomic MS repositories and pipelines. Proteomics 9(4):861-81*

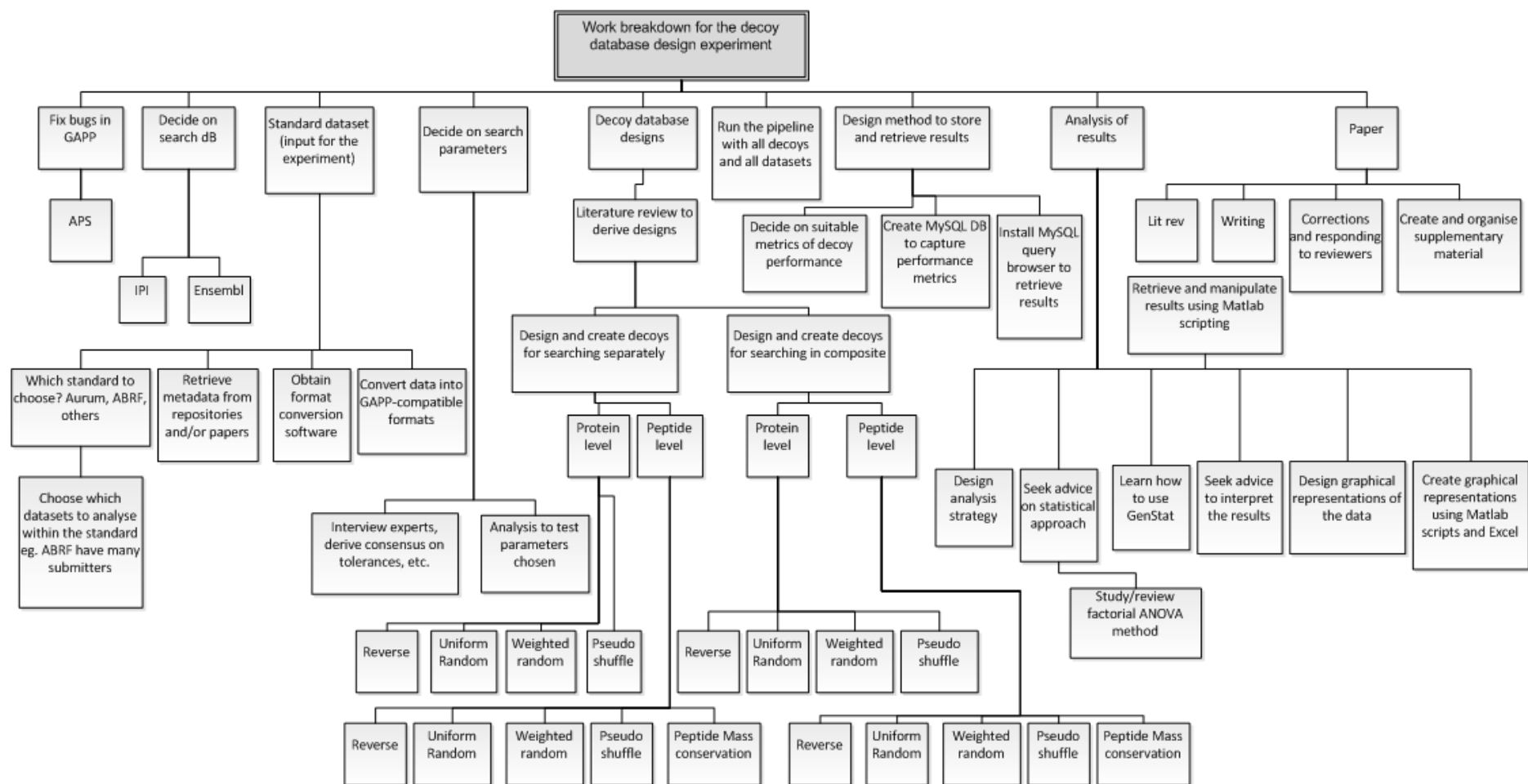
paper3_2009_proteomics.pdf

Appendix IV

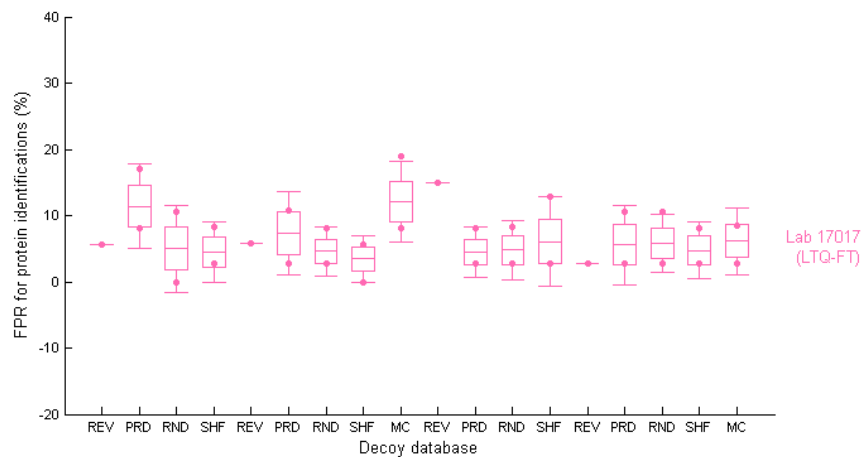
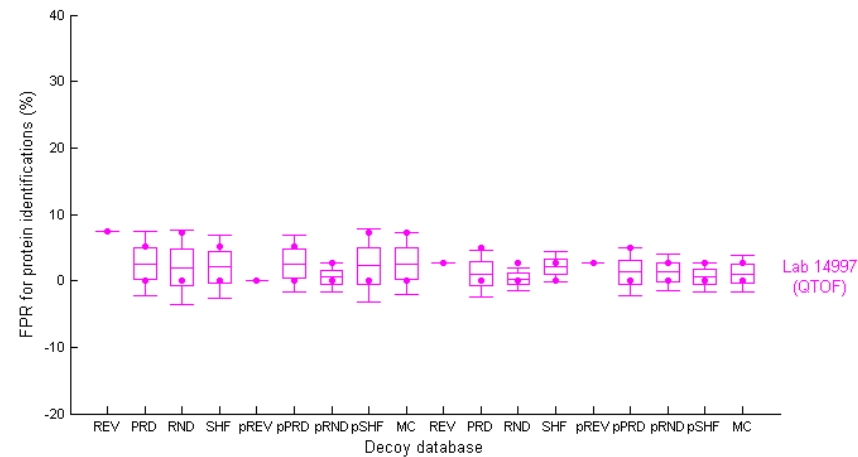
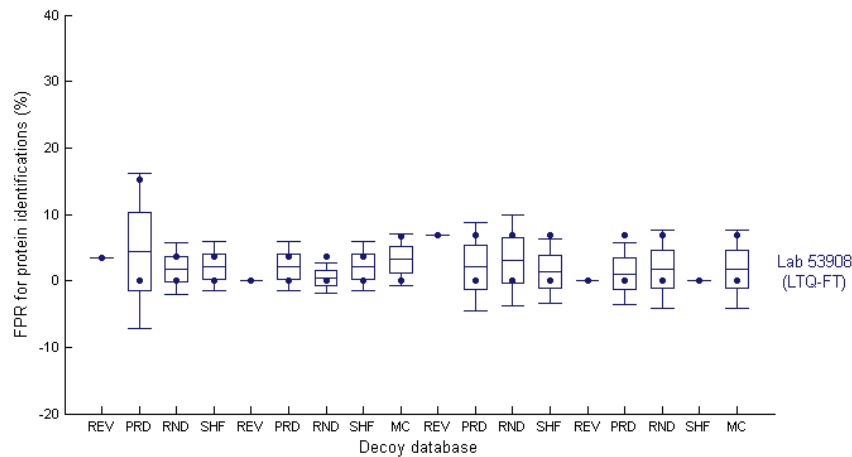
Paper 4 of the EngD

- Bianco, L., Mead J.A. and Bessant, C. (2009) Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS datasets Journal of Proteome Research 8(4):1782–1791

paper4_2009_jpr.pdf

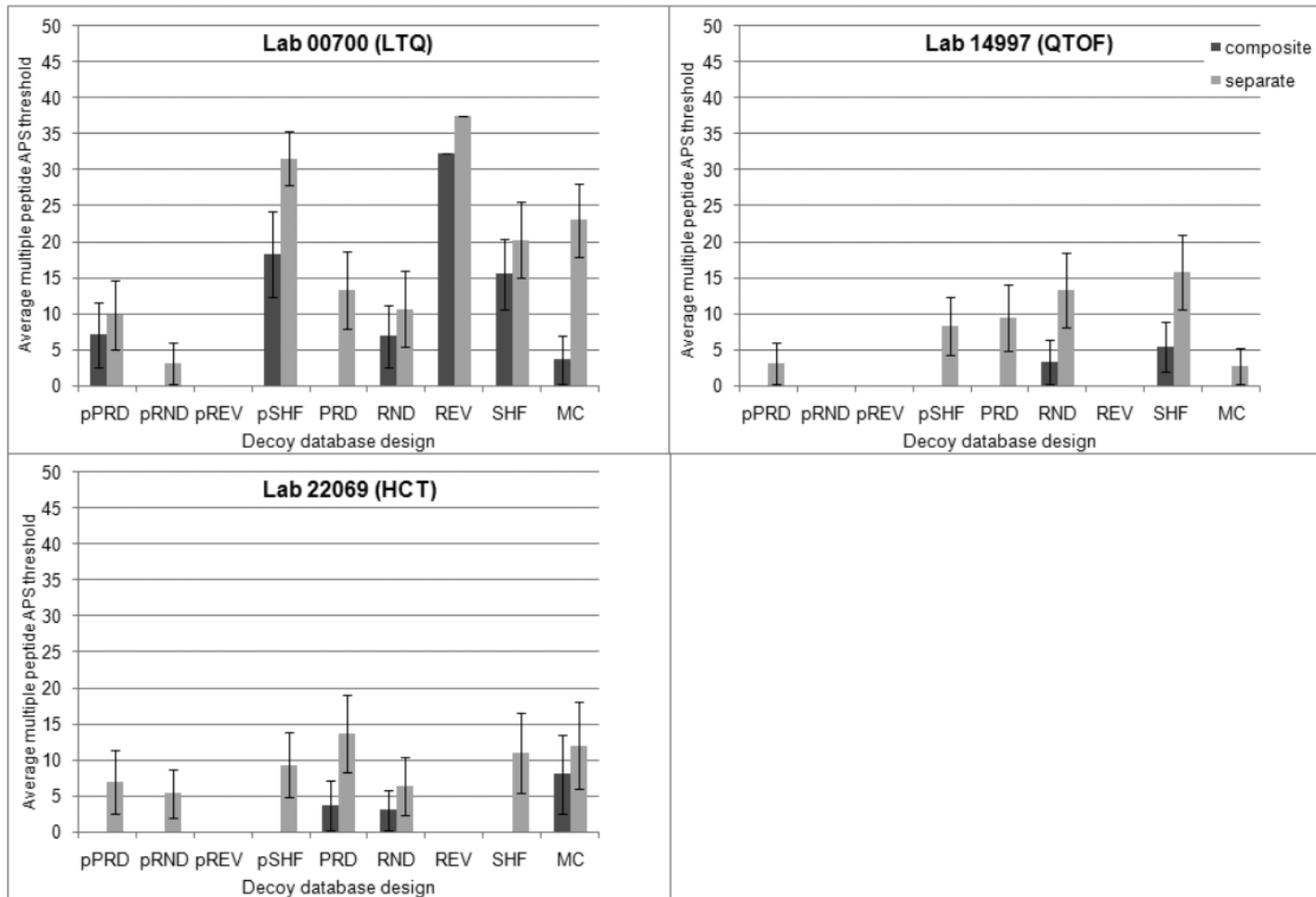


Appendix IV Figure 1 Work breakdown for the research work completed in Chapter 4

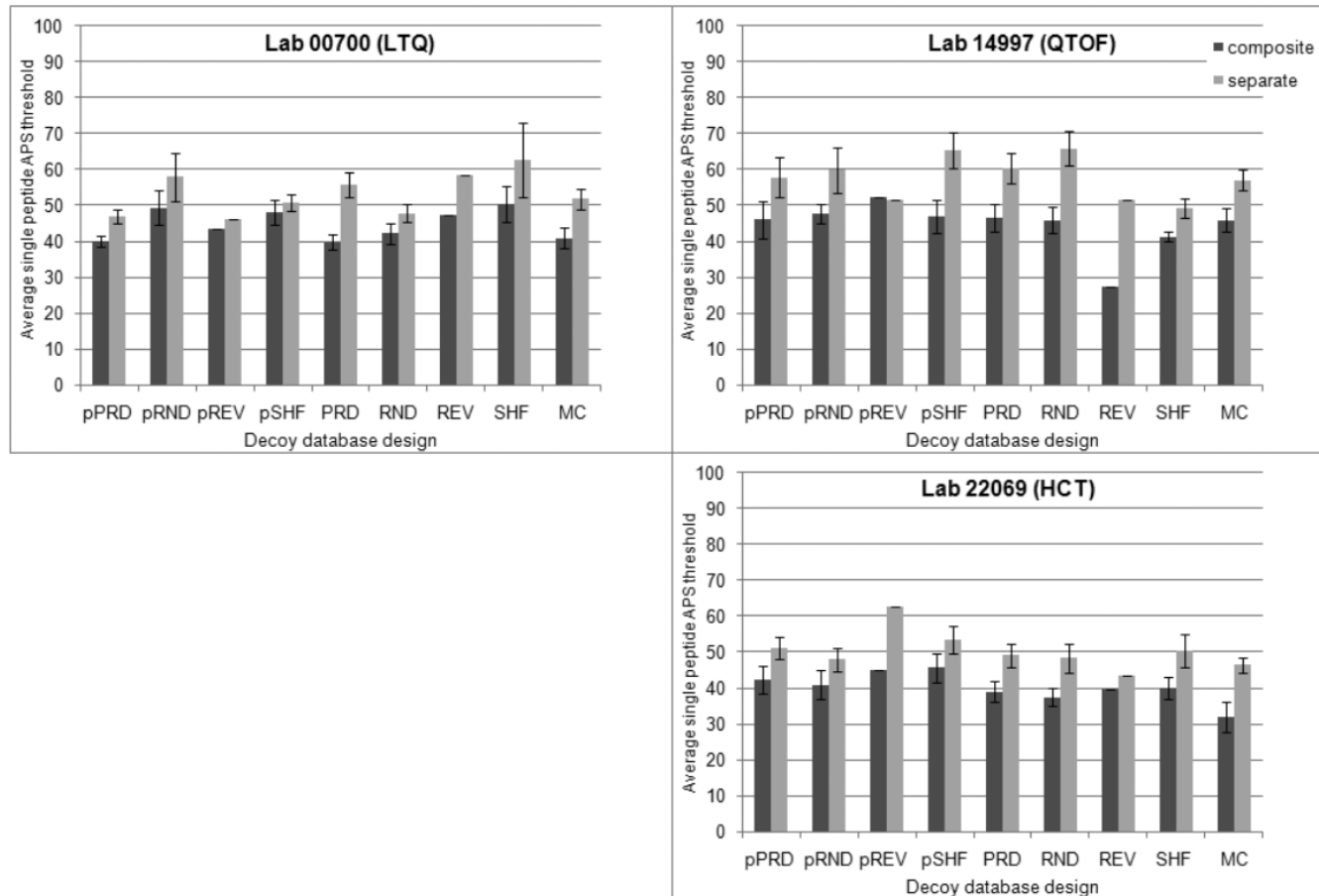


Appendix IV Figure 2 The remaining three labs from the box whisker plots in Chapter 4. The plot illustrate the distribution of protein level false positive rates (FPRs) across the ten database instances for each decoy design. Each colour represents an individual ABRF data-submitting laboratory. The horizontal line is the mean of the ten database instances, the box around the line shows one standard deviation above and below the mean, the whiskers show two standard deviations from the mean and the filled dots are the maximum and minimum individual FPR values obtained for the given decoy. The first nine decoys on the x-axis were searched in composite with the target and the last nine separate to the target.

(a)



(b)



Appendix IV Figure 3 Remaining graphs from Chapter 4.

Appendix IV Table 1 Examples of standard public MS/MS datasets and where to find them.

Dataset name	Reference article where available	Description	Instrument used	Data format	Data download location	Metadata location
ABRF sPRG2006	Paper pending see www.abrf.org	derived from a study involving anonymous analysis by multiple labs. 49 human proteins, plus human and non-human bonus protein list determined by the sPRG BIC2007 consensus study	Various including HCT, LCQ, LTQ, LTQ-FT, Q-TOF	.pkl .mgf native .raw	http://www.proteomecommons.org/data/show.jsp?id=802	http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm
ABRF iPRG2008	www.abrf.org /iprg has slides and posters	mouse liver differential expression using iTRAQ	Various	.mgf native.raw mzData mzXML .dta	http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm	http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm
Aurum	(Falkner <i>et al.</i> , 2007)	over 250 known human proteins	MALDI-TOF/TOF	.mgf .pkl .t2d	http://www.proteomecommons.org/data/show.jsp?id=90	'Aurum homepage' http://www.proteomecommons.org/archive/1122567790437/index.html
The ISB's standard protein mix (SPMDB)	(Klimek <i>et al.</i> , 2008)	18 proteins 1.1million spectra including 150 replicate runs. The standard is used regularly for determining performance of in-house MS instruments Species include rabbit, bovine, human, <i>B. licheniformis</i> , chicken	LTQ, LCQ Deca, Q-TOF, QSTAR, XCT Ultra, ABI 4800, ABI 4700, LTQ-FT	native .raw mzXML	http://regisweb.systemsbiology.net/PublicDatasets/ Very large dataset, hence suitable for training and validation	Most useful information is in the accompanying publication
SASHIMI 17 protein standard	None	17 tryptically digested proteins, multiple species including bovine, rabbit, <i>E.coli</i> , chicken, horse and others	Micromass Q-TOF Ultima	mzXML	http://sashimi.sourceforge.net/repository.html	http://sashimi.sourceforge.net/repository.html

SASHIMI 7 protein standard	None	7 protein mix: - rabbit glycogen phosphorylase, <i>E. Coli</i> beta- galactosidase, bovine serum albumin, myosin, chicken ovalbumin, bovine serotransferrin	LCQ	mzXML	http://sashimi.sourceforge.net/repository.html	http://sashimi.sourceforge.net/repository.html
SASHIMI 7 protein ICAT standard	None	Cleavable ICAT labelled 7 protein mix: bovine catalase, bovine alpha- lactalbumin, chicken ovalbumin, bovine serum albumin, horse myoglobin, bovine serotransferrin, rabbit glycogen phosphorylase	LCQ	mzXML	http://sashimi.sourceforge.net/repository.html	http://sashimi.sourceforge.net/repository.html
Experimen tal Protein Mixture	(Keller <i>et al.</i> , 2002b)	Tandem mass spectra for 14 LC/MS/MS runs of control mixture A and 8 LC/MS/MS runs on control mixture B (explained in paper)	Thermo ion trap	native .raw, .dta	http://www.systemsbiology.org/extra/protein_mixture.html	Data is downloaded from a password protected website (access can be granted for both non-commercial and commercial upon application by email)

Appendix IV Table 2 The identifiers considered to be true positives in this study performed in Chapter 4. They are derived from all the human proteins in the published ABRF sPRG 'BIC final protein list' (downloaded from: <http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm>). These identifiers were generated from SWISSPROT accession numbers using the PICR program at the EBI (<http://www.ebi.ac.uk/Tools/picr>).

ENSG00000012223
ENSG00000015475
ENSG00000017427
ENSG00000083750
ENSG00000084207
ENSG00000090013
ENSG00000090382
ENSG00000091513
ENSG00000091583
ENSG00000096087
ENSG00000096696
ENSG00000100311
ENSG00000100448
ENSG00000100665
ENSG00000102081
ENSG00000103275
ENSG00000104267
ENSG00000104879
ENSG00000105220
ENSG00000106804
ENSG00000109107
ENSG00000112855
ENSG00000116030
ENSG00000117450
ENSG00000117601
ENSG00000117984
ENSG00000119392
ENSG00000121691
ENSG00000121769
ENSG00000124588
ENSG00000125730
ENSG00000129559
ENSG00000132141
ENSG00000132693
ENSG00000133703
ENSG00000133742
ENSG00000134202
ENSG00000136810
ENSG00000138207
ENSG00000138798
ENSG00000139610
ENSG00000142168
ENSG00000143416
ENSG00000143437
ENSG00000143947
ENSG00000148180
ENSG00000149575
ENSG00000149925
ENSG00000150991
ENSG00000155876
ENSG00000163631
ENSG00000163815
ENSG00000164111
ENSG00000166347
ENSG00000166710

ENSG00000167244
ENSG00000167531
ENSG00000167768
ENSG00000167815
ENSG00000169429
ENSG00000170035
ENSG00000170142
ENSG00000170315
ENSG00000170442
ENSG00000170445
ENSG00000170465
ENSG00000170523
ENSG00000171345
ENSG00000171346
ENSG00000171401
ENSG00000171403
ENSG00000172115
ENSG00000172232
ENSG00000172379
ENSG00000172867
ENSG00000173636
ENSG00000173801
ENSG00000174156
ENSG00000174697
ENSG00000174775
ENSG00000175063
ENSG00000176919
ENSG00000181019
ENSG00000182247
ENSG00000182793
ENSG00000185479
ENSG00000186081
ENSG00000186395
ENSG00000186442
ENSG00000186831
ENSG00000186832
ENSG00000186847
ENSG00000186868
ENSG00000187681
ENSG00000188170
ENSG00000188536
ENSG00000196084
ENSG00000196262
ENSG00000196565
ENSG00000198125
ENSG00000198618
ENSG00000203786
ENSG00000204319
ENSG00000204490
ENSG00000205420
ENSG00000205426
ENSG00000206172
ENSG00000206328
ENSG00000206439
ENSG00000211592
ENSG00000211679
ENSG00000211890
ENSG00000211895
ENSG00000211896
ENSG00000211899
ENSG00000213281
ENSG00000213931

Appendix V

Paper 5 of the EngD

- Mead, J.A., Bianco, L., Ottone, V., Barton, C., Kay, R.G., Lilley, K.S., Bond, N. and Bessant, C. (2009) *MRMaid: the web-based tool for design of multiple reaction monitoring (MRM) transitions*, *Mol Cell Proteomics* 8(4): 696-705

paper5_2009_mcp.pdf

Paper 6 of the EngD

- Mead, J.A., Bianco, L. and Bessant, C. (2009) *Mining proteomic MS/MS data for MRM transitions*. *Methods in Mol. Biol.*;604:187-99

paper6_2009_humana.pdf

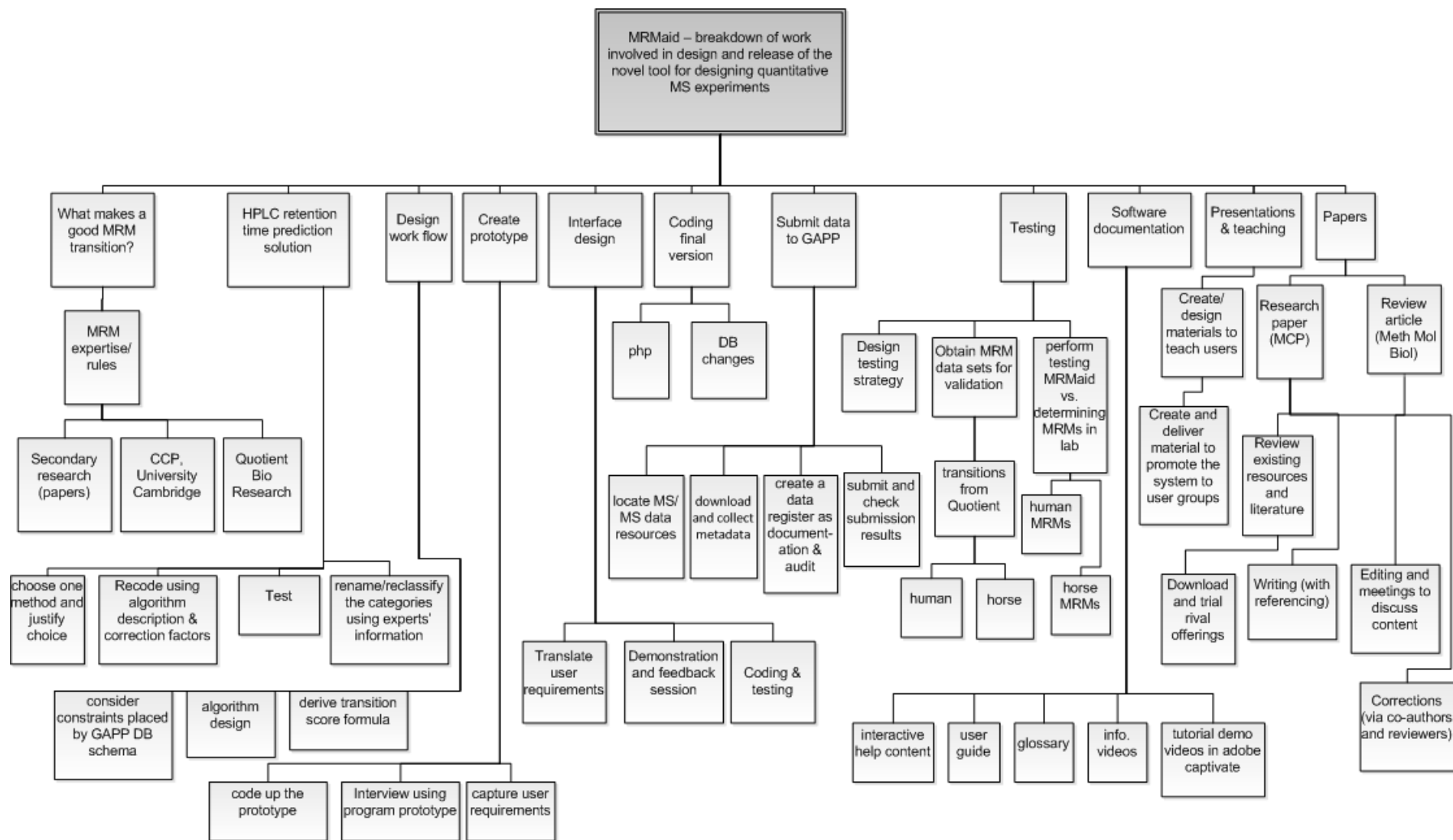
MRMaid user guide

mrmaid_userguide.pdf

MRMaid glossary

mrmaid_glossary.pdf

Appendix V Figure 1 Work breakdown for the research work completed in Chapter 5



Appendix VI

Paper 7 of the EngD

- Mead J.A., Bianco,L., Barton C. and Bessant,C (2010) MR Maid-DB: a compendium of published SRM transitions. *Journal of Proteome Research* 9(1):620-5

paper7_2010_jpr.pdf

MR Maid-DB user guide

mrmaid_db_userguide.pdf

Folder ('biomart') contains all the files needed for the set up of the Biomart instance for MR Maid-DB. Important files include:

Biomart registry file required to set up the MR Maid-DB Biomart query interface

myRegistryTransitionEns.xml

Data to populate the MRM MySQL database (16 tables)

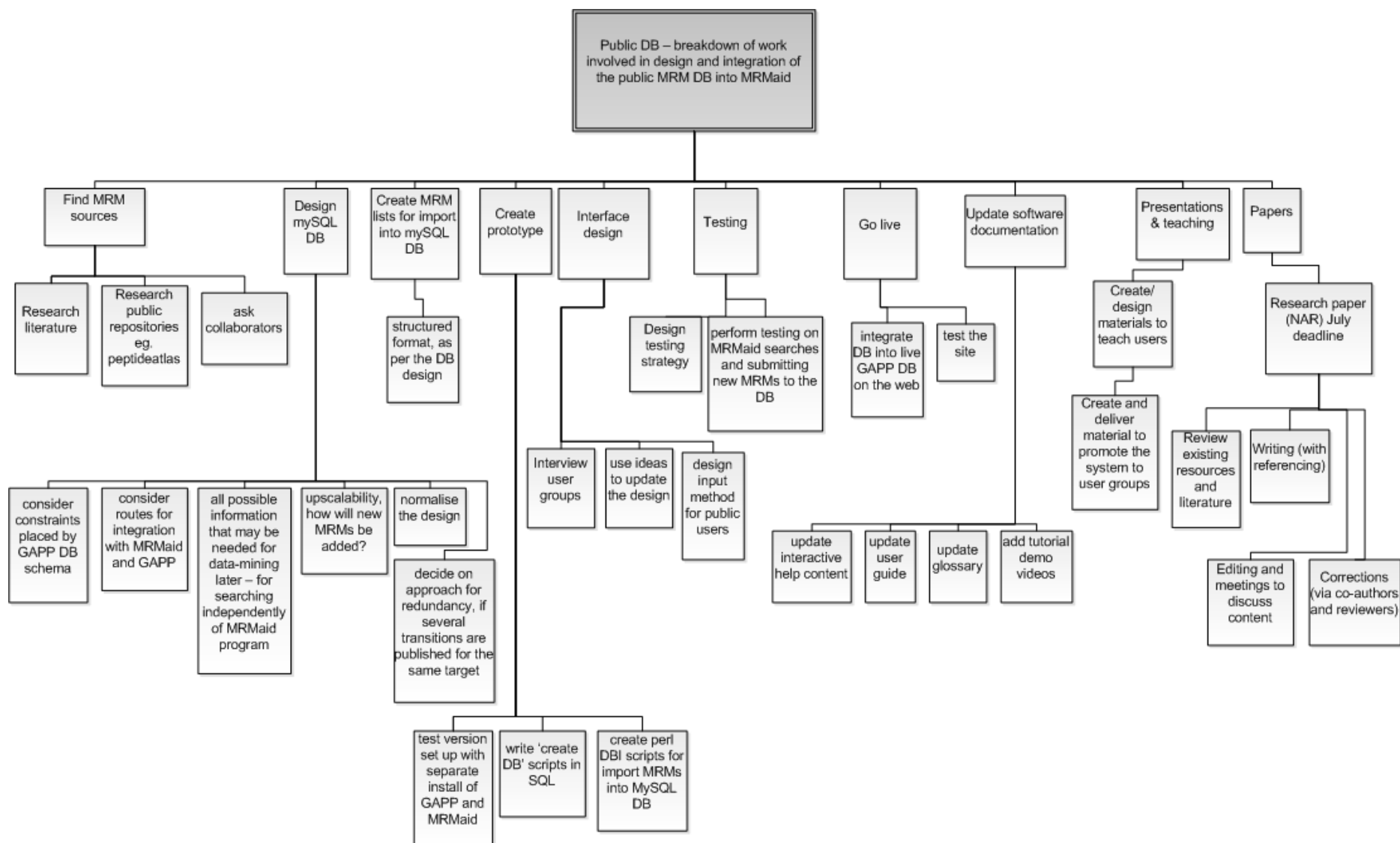
csv_mrmaid_db_data_population.txt

SQL dump file of the populated database

mrmaid_db_dump.sql

Apache server configuration file

httpd.conf



Appendix VI Figure 1 Work breakdown for the research work completed in Chapter 6

Appendix VII

Paper 8 of the EngD

- Mead, J.A., Bianco, L. and Bessant C. (2010) Free computational resources for designing selected reaction monitoring (SRM) transitions. *Proteomics* Jan 13th 2010. [Epub ahead of print]

paper8_2010_proteomics.pdf

