

Contents

Chapter 1	Introduction to Proteome Informatics	1
	<i>Conrad Bessant</i>	
1.1	Introduction	1
1.2	Principles of LC-MS/MS Proteomics	3
1.2.1	Protein Fundamentals	3
1.2.2	Shotgun Proteomics	5
1.2.3	Separation of Peptides by Chromatography	6
1.2.4	Mass Spectrometry	6
1.3	Identification of Peptides and Proteins	8
1.4	Protein Quantitation	9
1.5	Applications and Downstream Analysis	9
1.6	Proteomics Software	10
1.6.1	Proteomics Data Standards and Databases	11
1.7	Conclusions	12
	Acknowledgements	12
	References	12

Section I: Protein Identification

Chapter 2	<i>De novo</i> Peptide Sequencing	17
	<i>Bin Ma</i>	
2.1	Introduction	17
2.2	Manual <i>De novo</i> Sequencing	18
2.3	Computer Algorithms	20

New Developments in Mass Spectrometry No. 5

Proteome Informatics

Edited by Conrad Bessant

© The Royal Society of Chemistry 2017

Published by the Royal Society of Chemistry, www.rsc.org

2.3.1 Search Tree Pruning	20
2.3.2 Spectrum Graph	21
2.3.3 PEAKS Algorithm	24
2.4 Scoring Function	26
2.4.1 Likelihood Ratio	27
2.4.2 Utilization of Many Ion Types	28
2.4.3 Combined Use of Different Fragmentations	28
2.4.4 Machine Learning	29
2.4.5 Amino Acid Score	30
2.5 Computer Software	31
2.5.1 Lutefisk	31
2.5.2 Sherenga	31
2.5.3 PEAKS	31
2.5.4 PepNovo	32
2.5.5 DACSIM	32
2.5.6 NovoHMM	32
2.5.7 MSNovo	32
2.5.8 PILOT	32
2.5.9 pNovo	33
2.5.10 Novor	33
2.6 Conclusion: Applications and Limitations of <i>De novo</i> Sequencing	33
2.6.1 Sequencing Novel Peptides and Detecting Mutated Peptides	33
2.6.2 Assisting Database Search	34
2.6.3 <i>De novo</i> Protein Sequencing	34
2.6.4 Unspecified PTM Characterization	34
2.6.5 Limitations	35
Acknowledgements	35
References	36

Chapter 3 Peptide Spectrum Matching *via* Database Search and Spectral Library Search 39

Brian Netzel and Surendra Dasari

3.1 Introduction	39
3.2 Protein Sequence Databases	41
3.3 Overview of Shotgun Proteomics Method	43
3.4 Collision Induced Dissociation Fragments Peptides in Predictable Ways	44
3.5 Overview of Database Searching	45
3.6 MyriMatch Database Search Engine	47
3.6.1 Spectrum Preparation	48
3.6.2 Peptide Harvesting from Database	49
3.6.3 Comparing Experimental MS/MS with Candidate Peptide Sequences	49

3.7	Accounting for Post-Translational Modifications During Database Search	52
3.8	Reporting of Database Search Peptide Identifications	53
3.9	Spectral Library Search Concept	55
3.10	Peptide Spectral Libraries	56
3.11	Overview of Spectral Library Searching	58
3.12	Pepitome Spectral Library Search Engine	59
3.12.1	Experimental MS2 Spectrum Preparation	60
3.12.2	Library Spectrum Harvesting and Spectrum-Spectrum Matching	60
3.12.3	Results Reporting	62
3.13	Search Results Vary Between Various Database Search Engines and Different Peptide Identification Search Strategies	62
3.14	Conclusion	63
	References	64
Chapter 4	PSM Scoring and Validation	69
	<i>James C. Wright and Jyoti S. Choudhary</i>	
4.1	Introduction	69
4.2	Statistical Scores and What They Mean	71
4.2.1	Statistical Probability p -Values and Multiple Testing	72
4.2.2	Expectation Scores	72
4.2.3	False Discovery Rates	73
4.2.4	q -Values	74
4.2.5	Posterior Error Probability	75
4.2.6	Which Statistical Measure to Use and When	75
4.2.7	Target Decoy Approaches for FDR Assessment	77
4.3	Post-Search Validation Tools and Methods	80
4.3.1	Quality	80
4.3.2	PeptideProphet	81
4.3.3	Percolator	81
4.3.4	Mass Spectrometry Generating Function	82
4.3.5	Nokoi	83
4.3.6	PepDistiller	83
4.3.7	Integrated Workflow and Pipeline Analysis Tools	83
4.3.8	Developer Libraries	84
4.4	Common Pitfalls and Problems in Statistical Analysis of Proteomics Data	84
4.4.1	Target-Decoy Peptide Assumptions	84
4.4.2	Peptide Modifications	85
4.4.3	Search Space Size	86

4.4.4 Distinct Peptides and Proteins	87
4.5 Conclusion and Future Trends	88
References	88

Chapter 5 Protein Inference and Grouping 93

Andrew R. Jones

5.1 Background	93
5.1.1 Assignment of Peptides to Proteins	95
5.1.2 Protein Groups and Families	97
5.2 Theoretical Solutions and Protein Scoring	100
5.2.1 Protein Grouping Based on Sets of Peptides	100
5.2.2 Spectral-Focussed Inference Approaches	102
5.2.3 Considerations of Protein Length	104
5.2.4 Handling Sub-Set and Same-Set Proteins within Groups	105
5.2.5 Assignment of Representative or Group Leader Proteins	108
5.2.6 Importance of Peptide Classification to Quantitative Approaches	108
5.2.7 Scoring or Probability Assignment at the Protein-Level	109
5.2.8 Handling "One Hit Wonders"	111
5.3 Support for Protein Grouping in Data Standards	112
5.4 Conclusions	113
Acknowledgements	114
References	114

Chapter 6 Identification and Localization of Post-Translational Modifications by High-Resolution Mass Spectrometry 116

Rune Matthiesen and Ana Sofia Carvalho

6.1 Introduction	116
6.2 Sample Preparation Challenges	118
6.3 Identification and Localization of Post-Translational Modifications	120
6.3.1 Computational Challenges	120
6.3.2 Annotation of Modifications	122
6.3.3 Common Post-Translational Modifications Identified by Mass Spectrometry	123
6.3.4 Validation of Results	124
6.4 Conclusion	129
Acknowledgements	129
References	129

Section II: Protein Quantitation

Chapter 7	Algorithms for MS1-Based Quantitation	135
	<i>Hanqing Liao, Alexander Phillips, Andris Jankevics and Andrew W. Dowsey</i>	
7.1	Introduction	135
7.2	Feature Detection and Quantitation	137
7.2.1	Conventional Feature Detection	138
7.2.2	Recent Approaches Based on Sparsity and Mixture Modelling	140
7.3	Chromatogram Alignment	142
7.3.1	Feature-Based Pattern Matching	143
7.3.2	Raw Profile Alignment	143
7.4	Abundance Normalisation	146
7.5	Protein-Level Differential Quantification	147
7.5.1	Statistical Methods	148
7.5.2	Statistical Models Accounting for Shared Peptides	151
7.6	Discussion	151
	Acknowledgements	152
	References	152
Chapter 8	MS2-Based Quantitation	155
	<i>Marc Vaudel</i>	
8.1	MS2-Based Quantification of Proteins	155
8.2	Spectral Counting	156
8.2.1	Implementations	158
8.2.2	Conclusion on Spectrum Counting	158
8.3	Reporter Ion-Based Quantification	161
8.3.1	Identification	164
8.3.2	Reporter Ion Intensities, Interferences and Deisotoping	165
8.3.3	Ratio Estimation and Normalization	168
8.3.4	Implementation	169
8.3.5	Conclusion on Reporter Ion-Based Quantification	173
	Acknowledgements	175
	References	175
Chapter 9	Informatics Solutions for Selected Reaction Monitoring	178
	<i>Birgit Schilling, Brendan Maclean, Jason M. Held and Bradford W. Gibson</i>	
9.1	Introduction	178
9.1.1	SRM – General Concept and Specific Bioinformatic Challenges	178

9.1.2 SRM-Specific Bioinformatics Tools	180
9.2 SRM Assay Development	182
9.2.1 Target and Transition Selection, Proteotypic and Quantotypic Peptides	182
9.2.2 Spikes of Isotopically Labeled Peptides and Protein Standards and Additional Assay Development Steps	183
9.2.3 Retention Time Regressions and Retention Time Scheduling	184
9.2.4 Method Generation for MS Acquisitions	186
9.3 System Suitability Assessments	188
9.4 Post-Acquisition Processing and Data Analysis	188
9.4.1 mProphet False Discovery Analysis, Peak Detection and Peak Picking	188
9.4.2 Data Viewing and Data Management: Custom Annotation, Results and Document Grids, Group Comparisons	191
9.4.3 Data Reports, LOD-LOQ Calculations and Statistical Processing, Use of Skyline External Tools	191
9.4.4 Group Comparisons and Peptide & Protein Quantification	192
9.4.5 Easy Data Sharing and SRM Resources – Panorama	193
9.5 Post-Translational Modifications and Protein Isoforms or Proteoforms	193
9.6 Conclusion and Future Outlook	195
Acknowledgements	196
References	196
Chapter 10 Data Analysis for Data Independent Acquisition	200
<i>Pedro Navarro, Marco Trevisan-Herraz and Hannes L. Röst</i>	
10.1 Analytical Methods	200
10.1.1 Motivation	200
10.1.2 Background: Other MS Methods	201
10.1.3 DIA Concept	202
10.1.4 Theoretical Considerations	204
10.1.5 Main DIA Methods	207
10.1.6 Analyte Separation Methods	210
10.2 Data Analysis Methods	212
10.2.1 DIA Data Analysis	212
10.2.2 Untargeted Analysis, Spectrum-Centric	213
10.2.3 Targeted Analysis, Chromatogram-Centric	215
10.2.4 FDR	220
10.2.5 Results and Formats	222

10.3 Challenges	223
References	224

Section III: Open Source Software Environments for Proteome Informatics

Chapter 11	Data Formats of the Proteomics Standards Initiative	231
	<i>Juan Antonio Vizcaino, Simon Perkins, Andrew R. Jones and Eric W. Deutsch</i>	

11.1	Introduction	231
11.2	mzML	233
11.2.1	Data Format	233
11.2.2	Software Implementations	235
11.2.3	Current Work	237
11.2.4	Variations of mzML	237
11.3	mzIdentML	238
11.3.1	Data Format	238
11.3.2	Software Implementations	241
11.3.3	Current Work	242
11.4	mzQuantML	242
11.4.1	Data Format	242
11.4.2	Software Implementations	245
11.4.3	Current Work	245
11.5	mzTab	246
11.5.1	Data Format	246
11.5.2	Software Implementations	248
11.5.3	Current Work	249
11.6	TraML	249
11.6.1	Data Format	249
11.6.2	Software Implementations	251
11.7	Other Data Standard Formats Produced by the PSI	251
11.8	Conclusions	252
	Abbreviations	252
	Acknowledgements	253
	References	253

Chapter 12	OpenMS: A Modular, Open-Source Workflow System for the Analysis of Quantitative Proteomics Data	259
	<i>Lars Nilse</i>	

12.1	Introduction	259
12.2	Peptide Identification	262
12.3	iTRAQ Labeling	266
12.4	Dimethyl Labeling	270

12.5 Label-Free Quantification	275
12.6 Conclusion	279
Acknowledgements	282
References	282
Chapter 13 Using Galaxy for Proteomics	289
<i>Candace R. Guerrero, Pratik D. Jagtap, James E. Johnson and Timothy J. Griffin</i>	
13.1 Introduction	289
13.2 The Galaxy Framework as a Solution for MS-Based Proteomic Informatics	291
13.2.1 The Web-Based User Interface	291
13.2.2 Galaxy Histories	293
13.2.3 Galaxy Workflows	293
13.2.4 Sharing Histories and Workflows in Galaxy	296
13.3 Extending Galaxy for New Data Analysis Applications	296
13.3.1 Deploying Software as a Galaxy Tool	296
13.3.2 Galaxy Plugins and Visualization	299
13.4 Publishing Galaxy Extensions	300
13.5 Scaling Galaxy for Operation on High Performance Systems	300
13.6 Windows-Only Applications in a Linux World	301
13.7 MS-Based Proteomic Applications in Galaxy	302
13.7.1 Raw Data Conversion and Pre-Processing	302
13.7.2 Generation of a Reference Protein Sequence Database	304
13.7.3 Sequence Database Searching	304
13.7.4 Results Filtering and Visualization	305
13.8 Integrating the 'omic' Domains: Multi-Omic Applications in Galaxy	306
13.8.1 Building Proteogenomic Workflows in Galaxy	309
13.8.2 Metaproteomics Applications in Galaxy	313
13.9 Concluding Thoughts and Future Directions	315
Acknowledgements	317
References	317
Chapter 14 R for Proteomics	321
<i>Lisa M. Breckels, Sebastian Gibb, Vladislav Petyuk and Laurent Gatto</i>	
14.1 Introduction	321
14.2 Accessing Data	323
14.2.1 Data Packages	323

14.2.2	Data from the ProteomeXchange Repository	324
14.2.3	Cloud Infrastructure	325
14.3	Reading and Handling Mass Spectrometry and Proteomics Data	326
14.3.1	Raw Data	326
14.3.2	Identification Data	327
14.3.3	Quantitative Data	329
14.3.4	Imaging Data	330
14.3.5	Conclusion	330
14.4	MSMS Identifications	330
14.4.1	Introduction	330
14.4.2	The MSGFplus Package	331
14.4.3	The MSGFgui Package	332
14.4.4	The rTANDEM Package	334
14.4.5	The MSnID Package	335
14.4.6	Example	338
14.5	Analysis of Spectral Counting Data	339
14.5.1	Introduction	339
14.5.2	Exploratory Data Analysis with msmsEDA	339
14.5.3	Statistical Analyses with msmsTests	341
14.5.4	Example	342
14.6	MALDI and Mass Spectrometry Imaging	342
14.6.1	Introduction	342
14.6.2	MALDI Pre-Processing Using MALDIquant	343
14.6.3	Mass Spectrometry Imaging	348
14.7	Isobaric Tagging and Quantitative Data Processing	350
14.7.1	Quantification of Isobaric Data Experiments	351
14.7.2	Processing Quantitative Proteomics Data	351
14.8	Machine Learning, Statistics and Applications	352
14.8.1	Introduction	352
14.8.2	Statistics	352
14.8.3	Machine Learning	354
14.8.4	Conclusion	358
14.9	Conclusions	359
	References	359

Section IV: Integration of Proteomics and Other Data

Chapter 15	Proteogenomics: Proteomics for Genome Annotation	367
	<i>Fawaz Ghali and Andrew R. Jones</i>	
15.1	Introduction	367
15.2	Theoretical Underpinning	370
15.2.1	Gene Prediction	371
15.2.2	Protein and Peptide Identification	372

15.2.3	Design of Protein Sequence Databases	372
15.2.4	Output of Proteogenomics Pipelines	375
15.3	Proteogenomics Platforms	377
15.3.1	Gene Prediction Pipelines	377
15.3.2	Proteogenomics Pipelines	378
15.3.3	Proteomics Data Repositories for Proteogenomics	378
15.3.4	Visualisation	379
15.3.5	Data Formats and Standards	380
15.4	Challenges and Future Research	381
15.5	Summary	381
	References	382
Chapter 16	Proteomics Informed by Transcriptomics	385
	<i>Shyamasree Saha, David Matthews and Conrad Bessant</i>	
16.1	Introduction to PIT	385
16.2	Creation of Protein Database from RNA-Seq Data	388
16.2.1	Introduction to RNA-Seq	388
16.2.2	Sequence Assembly	391
16.2.3	ORF Finding	392
16.2.4	Finalising Protein Sequence Data for PIT Search	393
16.3	Interpretation of Identified ORFs	393
16.3.1	Identification of Proteins in the Absence of a Reference Genome	394
16.3.2	Identification of Individual Sequence Variation	394
16.3.3	Monitoring Isoform Switching	397
16.3.4	Genome Annotation and Discovery of Novel Translated Genomic Elements	400
16.4	Reporting and Storing PIT Results	400
16.5	Applications of PIT	401
16.6	Conclusions	402
	Acknowledgements	402
	References	402
	Subject Index	406