PROTEIN IDENTIFICATION AND PROTEIN EXPRESSION PROFILING OF *SACCHAROMYCES CEREVISIAE* GROWN UNDER LOW AND VERY HIGH GRAVITY CONDITIONS

A Thesis

Submitted to the College of Graduate Studies and Research

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the Department of Chemical Engineering

University of Saskatchewan

Saskatoon, Saskatchewan Canada,

S7N 5C5

By

YUPENG ZHAO

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work, or in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Chemical Engineering

University of Saskatchewan

Research Annex, 105 Maintenance Road

Saskatoon, SK S7N 5C5 Canada

ABSTRACT

Proteomics is the analysis of the total complement of proteins expressed by a cell or organism grown under a specified condition. The obtained protein profile would provide a better understanding of phenotypic characteristics of a cell grown under pre-determined conditions. Mass spectrometric-based protein analysis is currently the standard method in proteomic studies; however, there are many limitations associated with its application. The major objectives of this study included the development of a strategy to analyze the confidence of identified proteins and the development of an algorithm to interpret the experimentally obtained mass spectral data.

A two-step strategy was developed to analyze the confidence of identified proteins. In the first step, the proteins identified by a single protein identification tool were classified into two groups: high confidence proteins that were identified by unique peptides, and low confidence proteins that were identified by non-unique peptides. In the second step, the proteins identified by different tools (e.g., SEQUEST and Mascot in our work) were cross-compared. After integrating the two-step analysis, the identified proteins were classified into four levels of confidence. The proteins that were identified by the presence of unique peptides and that were commonly identified by different tools were grouped into the highest confidence level - Level 4. Even though the number of proteins

in Level 4 was reduced significantly, the conclusions drawn from the proteins were more reliable.

According to the operation of tandem mass spectrometry and the characteristics of the peptides generated by site-specific protease digestion, a two-pass approach for identifying the species-specific proteins was developed. The approach can find all possible peptides corresponding to a precursor ion and gives detailed matching information of each peptide candidate to the experimental product ion series. According to the total number of matched product ions, the total number of matched b- and y- ions, and the contiguity characteristic of identified product ions, the peptide candidates were ranked decreasingly from the most probable to the least. Combined with the concept of unique peptide, the obtained most probable peptide can then be used to predict proteins existing in the original sample.

The developed two-pass approach and two-step strategy were then used to study the protein profiling of *Saccharomyces cerevisiae* cultivated in various gravity conditions (10 and 300 g glucose/l) in order to investigate the changes in central metabolic pathways of *S. cerevisiae*. Our fermentation data indicated that the higher glucose contents would result in lower cell growth and higher ethanol production (e.g., high ethanol concentration in fermentation broth). However, the relative ethanol yield as related to the glucose consumption was lower under higher glucose concentrations. The protein profile showed that a higher flux of nutrient was channelled into the pentose phosphate pathway when *S. cerevisiae* was grown under a high glucose concentration. The reason for this phenomenon might be that the cell needs more reducing power (e.g.,

NADPH) for the synthesis of macromolecules such as proteins, nucleic acids, and lipids. These materials are essential to the cell in order to modify its structure (e.g., cell wall), to survive osmotic stress and to replicate.

ACKNOWLEDGEMENTS

I would like to express my appreciation to my supervisor Dr. Yen-Han Lin for his keen guidance, encouragement, and financial support during my Ph.D. program and his critical reviews and comments on my publications and this dissertation. I would also like to thank him giving me many opportunities of participating academic conferences.

I would like to thank my respected advisory committee members: Drs John V. Headley, Gordon A. Hill, W. Mike Ingledew, Darren R. Korber, and William J. Roesler. I consider myself fortunate to have had their invaluable suggestions during the course of my Ph.D. studies. Without their kind help, this thesis would not be done.

I am also grateful to Mr. Doug Olson, Mr. Ken Thoms, and Mr Kerry Peru, for their knowledge, experience and assistances in mass spectrometry analysis.

Finally, I would like to thank my family for their love and support. Particularly, I would like to thank my wife, Wanling Pan, for her continuous inspiration and my son, Peter, for bringing me a lot of happiness during my study.

TABLE OF CONTENTS

PERMISSION TO USE i
ABSTRACT ii
ACKNOWLEDGEMENTS v
TABLE OF CONTENTS vi
LIST OF FIGURES ix
LIST OF TABLES xi
LIST OF ABBREVIATIONS xii
LIST OF AMINO ACIDS AND THEIR MASS IN RESIDUE STATE xiv

Chapter 1 Introduction 1
1.1 Background 1
1.2 Contributions 4
1.2.1 Main contributions 4
1.2.2 Other contributions 5
1.3 Thesis organization 6
1.4 References 7
Chapter 2 Proteomics and applications 8
2.1 Two-dimensional polyacrylamide gel electrophoresis 8
2.2 Mass spectrometry for protein analysis 12
2.3 Multi-dimensional protein identification technology 18
2.4 Protein identification tools 21
2.5 Applications of proteomics 27
2.5.1 Protein profiling 28
2.5.2 Protein quantification 29
2.5.3 Mapping of protein modification 32
2.5.4 Protein-network mapping 37
2.6 References 39
Chapter 3 Strategy for the confidence analysis of identified proteins 48
3.1 Introduction 49
3.1.1 Protein identification using MS-Tag 49
3.1.2 Protein identification using Mascot 52
3.1.3 Protein identification using SEQUEST 59
3.1.4 Commentary for protein identification tools 61

3.2 Strategy for the confidence analysis of identified proteins 68
3.2.1 Unique peptide 70
3.2.2 Cross comparison 75
3.3 Implementation 76
3.4 Concluding remarks 80
3.5 References 80
Chapter 4 A proteomic tool for protein identification from tandem mass spectral data 82
4.1 Abstract 82
4.2 Introduction 83
4.3 Methods 84
4.3.1 Logic 84
4.3.2 Ranking criteria 86
4.4 Implementation 87
4.5 Concluding remarks 91
4.6 References 92
Chapter 5 Case studies: growth of *Saccharomyces cerevisiae* under low and high glucose conditions 93
5.1 Growth of *Saccharomyces cerevisiae* in a chemostat under high glucose conditions 93
5.1.1 Abstract 94
5.1.2 Introduction 94
5.1.3 Materials and methods 96
5.1.4 Results and discussion 98
5.1.5 Concluding remarks 103
5.1.6 Acknowledgments 104
5.1.7 References 105
5.1.8 Additional experimental details 106
5.2 A proteomic study of *Saccharomyces cerevisiae* grown under high specific gravity conditions 109
5.2.1 Abstract 109
5.2.2 Introduction 111
5.2.3 Materials and methods 112
5.2.4 Results and discussion 113
5.2.5 Concluding remarks 123
5.2.6 References 124
5.2.7 Additional experimental details 125
Chapter 6 Conclusions and future work 132
6.1 Generation discussion 132
6.2 Conclusions 134
6.3 Future work 135
6.3.1 Development of protein or peptide enrichment techniques 135
6.3.2 Improvement of our developed tools 135

Appendix A The development of an algorithm for the mass spectral interpretation of phosphoproteins 136
A.1 Abstract 136
A.2 Introduction 136
A.3 Methods 138
A.4 Implementation 139
A.5 Concluding remarks 144
A.6 References 145
Appendix B An automated approach to extract metabolically related proteins for metabolic flux analysis of *Pseudomonas putida* 146
B.1 Abstract 146
B.2 Introduction 147
B.3 Strategy and implementation 149
B.3.1 Gene database and Step 1 151
B.3.2 Reaction database and Step 2 154
B.3.3 Pathway database and Step 3 157
B.3.4 Integration output from above steps (Step 4) 159
B.4 Discussion and remarks 161
B.5 References 164
Appendix C Supplementary data for the developed two-pass approach 165
Appendix D Supplementary data for phosphopeptide identification 193

LIST OF FIGURES

Figure 2.1 Fundamental components of a mass spectrometer	13
Figure 2.2 Schematic of peptide fragmentation (http://www.matrixscience.com/help/fragmentation_help.html)	16
Figure 2.3 Partial listing of Protein P00549 and its peptides	35
Figure 3.1 Partial listing of an MS-Tag report	51
Figure 3.2 Partial listing of a Mascot protein summary report	53
Figure 3.3 Partial listing of a Mascot peptide summary report	54
Figure 3.4 Partial listing of a Mascot un-assigned peptide summary report	55
Figure 3.5 Sequence coverage for protein identification	57
Figure 3.6 Proteins predicted from just one peptide	58
Figure 3.7 Partial listing of a SEQUEST report	60
Figure 3.8 Partial listing of an MS-Tag summary report	62
Figure 3.9 Protein predicted from peptides with low probability	65
Figure 3.10 Several proteins predicted from the same identified peptide(s)	67
Figure 3.11 Schematic of strategy to analyze protein identification confidence	69
Figure 3.12 Amino acid sequences of pseudo proteins 1 and 2	71
Figure 3.13 Distributions of unique peptides, peptides, and protein candidates	74
Figure 3.14 Proteins identified with various degree of confidence using the proposed two-step strategy	77
Figure 4.1 General strategies for peptide and protein identification	85
Figure 5.1 Specific consumption (SGCR) and/or production (SEPR) rates of glucose and ethanol by *S. cerevisia*e grown in a chemostat under various glucose concentrations	100

Figure 5.2 Ethanol production yield coefficient of *S. cerevisiae* grown in a chemostat under various glucose concentrations 102

Figure 5.3 Schematics of a chemostat fermentation system 108

Figure 5.4 Schematics of central metabolic pathways used in *S. cerevisiae* 115

Figure 5.5 Partial listing of a PKL file (Fraction 4, 300 g glucose/l) 130

Figure A.1 (a) Illustrated search results for the precursor ion at m/z = 861.6 and z = 3, (b) Listing of all possible combinations of the phosphopeptide at m/z = 861.1 and z =3, (c) Listing of the identified most probable phosphopeptide at m/z = 861.6 and z =3. 143

Figure B.1 Schematic of proposed automated approach 150

Figure B.2 A partial listing of a gene database for *P. putida* 152

Figure B.3 A partial listing after Step 1 153

Figure B.4 A partial listing of a reaction database defined in KEGG 155

Figure B.5 A partial listing after Step 2 156

Figure B.6 A partial listing after Step 3 158

Figure B.7 A partial listing after Step 4 160

Figure B.8 A partial listing of the inter-relationship among genes, enzymes, and pathways 163

LIST OF TABLES

Table 2.1 Software tools for MS spectra interpretation	24
Table 4.1 Experimental MS spectral data extracted from Peng *et al*. (2003) and Prince *et al*. (2004)	88
Table 4.2 Searched results among several protein identification tools	90
Table 5.1 Concentrations of residual glucose and ethanol under various glucose concentrations	99
Table 5.2 Identified enzymes in the central metabolic pathway of *S. cerevisiae* grown at various glucose concentrations	114
Table 5.3 Concentrations of alanine and proline under various glucose concentrations	122
Table A.1 Pooled peptide database	140
Table A.2 Experimental MS spectral data retrieved from Synder (2000) and Wu *et al*.(2003)	142
Table C.1 Derived experimental data (known sequence)	166
Table C.2 List of matched peptide candidates (known sequence)	168
Table C.3 List of searched possible peptides (known sequence)	175
Table C.4 Derived experimental data (unknown sequence)	178
Table C.5 List of matched peptide candidates (unknown sequence)	180
Table C.6 List of searched possible peptides (unknown sequence)	188

LIST OF ABBREVIATIONS

2D-PAGE	Two-dimensional polyacrylamide gel electrophoresis
CID	Collision-induced dissociation
ESI-FTICR	Electrospray ionization fourier transform ion cyclotron resonance mass spectrometer
ESI-QIT	Electrospray ionization quadrupole ion trap mass spectrometer
ESI-Q-Q-Q	Electrospray ionization triple quadrupole mass spectrometer
ESI-Q-TOF	Electrospray ionization quadrupole time-of-flight mass spectrometer
ESI-TOF	Electrospray ionization time-of-flight mass spectrometer
ICAT	Isotope-coded affinity tag
IEF	Isoelectric focusing
IMAC	Immobilized metal affinity chromatography
IPG	Immobilized pH gradient
LC-LC	Two-dimensional liquid chromatography
m/z	Mass-to-charge ratio
MALDI-Q-TOF	Matrix-assisted laser desorption ionization quadrupole time-of-flight mass spectrometer
MALDI-TOF	Matrix-assisted laser desorption ionization time-of-flight mass spectrometer
MALDI-TOF-TOF	Matrix-assisted laser desorption ionization time-of-flight time-of-flight mass spectrometer
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry

MudPIT	Multi-dimensional protein identification technology
MW	Molecular weight
PFF	Peptide fragmentation fingerprinting
pI	Isoelectric point
PMF	Peptide mass fingerprinting
PTM	Post-translational modification
RP	Reverse phase
SCX	Strong cation exchange
SDS	Sodium dodecyl sulphate
TAP	Tandem affinity purification

LIST OF AMINO ACIDS AND THEIR MASS IN RESIDUE STATE

Amino acid	Codes		Mass (Da) of residue state[1]		
	3-letter	1-letter	Monoisotopic	Average	Nominal[2]
Alanine	Ala	A	71.03711	71.0788	71
Arginine	Arg	R	156.10111	156.1876	156
Asparagine	Asn	N	114.04293	114.1039	114
Aspartic acid	Asp	D	115.02649	115.0886	115
Cysteine	Cys	C	103.00919	103.1448	103
Glutamic acid	Glu	E	129.04259	129.1155	129
Glutamine	Gln	Q	128.05858	128.1308	128
Glycine	Gly	G	57.02146	57.0520	57
Histidine	His	H	137.05891	137.1412	137
Isoleucine	Ile	I	113.08406	113.1595	113
Leucine	Leu	L	113.08406	113.1595	113
Lysine	Lys	K	128.09496	128.1742	128
Methionine	Met	M	131.04049	131.1986	131
Phenylalanine	Phe	F	147.06841	147.1766	147
Proline	Pro	P	97.05276	97.1167	97
Serine	Ser	S	87.03203	87.0782	87
Threonine	Thr	T	101.04768	101.1051	101
Tryptophan	Trp	W	186.07931	186.2133	186
Tyrosine	Tyr	Y	163.06333	163.1760	163
Valine	Val	V	99.06841	99.1326	99

[1] The residue state of an amino acid refers to a state of the amino acid that has one H_2O missing (http://i-mass.com/guide/aamass.html).

[2] The nominal mass of a residue amino acid is the whole-number portion of the corresponding mass of the amino acid.

Chapter 1 Introduction

1.1 Background

Proteomic study is the global analysis of complex protein mixtures for the purpose of qualitative, quantitative and functional analysis of all the proteins present in a given cell, tissue or organism (Hunter *et al*., 2002). Proteins perform most of the metabolic and structural functions essential for the cell; therefore, the systematic analysis of proteins is necessary for a better understanding of cellular growth, development, replication, and stress response. Currently, most proteomic projects are grouped into four major subcategories: 1) identification and comparison of protein profiling in normal and abnormal cells; 2) quantification of proteins in a cell or organism; 3) characterization of proteins with post-translational modifications (PTM); and 4) mapping of protein-protein interactions. Correspondingly, the major applications of proteomics include: 1) profiling comparison of proteins; 2) quantification of proteins; 3) mapping of PTM proteins; and 4) investigation of protein-protein interactions.

Typically, proteomic analysis consists of a partition step that separates proteins or peptides from a complex protein mixture, and an analytical step that identifies and/or quantifies the expressed proteins. Proteins expressed in a cell have a wide range of variety in terms of physiochemical characteristics (e.g., size, molecular weight, charge,

hydrophobicity, and so on) due to the various structures and properties of the amino acid components. Many classic separation methods (e.g., size exclusion chromatography, centrifugation, ion exchange chromatography, affinity chromatography, reversed-phase liquid chromatography, and gel electrophoresis) are typically applied to separate proteins or peptides from complex mixtures. In practice, a number of proteins in the sample mixtures may have close or similar physiochemical properties, so that they might be co-eluted when only one separation technique is implemented, bringing difficulty for subsequent protein analysis. Alternatively, multiple separation steps, in which protein mixtures are separated several times on the basis of different physicochemical properties, are often required to completely separate protein mixture for proteomic studies. Ideally, a method that can separate as many proteins as possible in the fewest possible steps is desired. Currently, two dimensional polyacrylamide gel electrophoresis (2D-PAGE) and two dimensional liquid chromatography (LC-LC) are the two most widely used techniques in proteomic studies. The 2D-PAGE technique is normally used to separate intact proteins from the original protein mixture, whereas LC-LC separates peptide mixtures generated by proteolytically digesting the original protein mixtures with a site-specific protease (e.g., trypsin).

The separated proteins or peptides can then be identified by visualization (e.g., stained by chemical dyes) or by mass spectrometric methods. The mass spectrometric technique plays a more important role in proteomic studies, because it has many advantages over visualization. For example, a mass spectrometer has a wide dynamic detection range, is able to analyze multiple proteins in a single injection, and is capable of providing an accurate mass spectrum for protein identification at high confidence. In fact, the successful

introduction of mass spectrometry (MS) into biological analysis and the rapid development of MS design made 'real' proteomics research possible in the mid-1990s and the field is expanding rapidly (Hunter *et al*., 2002). Some examples of mass spectrometers with various performances for proteomics research include electrospray ionization coupled with single mass spectrometers (e.g., ESI quadrupole ion trap mass spectrometers, ESI-QIT; ESI time-of-flight mass spectrometers, ESI-TOF; and ESI fourier transform ion cyclotron resonance, ESI-FTICR), ESI coupled with tandem mass spectrometers (e.g., ESI quadrupole TOF, ESI-Q-TOF; ESI triple quadrupole mass spectrometers, ESI-Q-Q-Q), matrix-assisted laser desorption ionization coupled with single mass spectrometers (e.g., MALDI-TOF), and MALDI coupled with tandem mass spectrometers (e.g., MALDI-Q-TOF, MALDI-TOF-TOF). Several books (Dass, 2001; Hoffmann, 2002; Kinter and Sherman, 2000) and reviews (Yarmush and Jayaraman, 2002; Yates, 2004) are recommended for interested readers. The technique of LC-LC separation coupled with tandem mass spectrometry (MS/MS) analysis is referred to as multi-dimensional protein identification technology (MudPIT).

The main problem in mass spectrometric-based proteomic studies is the interpretation of mass spectra for protein identification. Several algorithms have been developed to automate the interpretation process. However, these protein identification tools often report different results for the same set of mass spectral data due to the different logic in various tools. The verified searched results from different tools may confuse biological researchers, and even seriously affect their conclusions and future plans. Therefore, a suitable method to analyze the results is necessary.

In this dissertation, a strategy was developed to classify the confidence level of identified proteins on the basis of the specific characteristic of unique peptides. The strategy was validated using the searched results by Mascot (Perkins *et al*., 1999) and SEQUEST (Eng *et al*., 1994; Yates *et al*., 1995), two widely used commercial packages used for publicly accessible MS spectral data. In addition, we also developed a two-pass algorithm to interpret the experimental MS spectral data. This algorithm was validated by comparing the searched results with those identified by other available protein identification tools using the same mass spectral data. Finally, the strategy and algorithm were used to compare the protein profiles of *Saccharomyces cerevisiae* grown at low and high glucose concentrations.

1.2 Contributions

This dissertation presents research on techniques to identify proteins from MS and MS/MS spectral data and the strategy to analyze the confidences of identified proteins and/or to locate the proteins with the highest identification confidence. Large portions of this dissertation have been published recently (Zhao and Lin, 2003, 2004, 2005a, 2005b).

1.2.1 Main contributions

- A strategy to analyze the confidences of identified proteins or to locate proteins with the highest confidence was developed in our laboratory (Chapter 3). The significant discrepancy between proteins identified by Mascot and SEQUEST raises general questions about proteomic analyses, such as: 1) what is the level of confidence of

these identified proteins? 2) how to apply the identified proteins in discussing a biological phenomenon? 3) should only one standardized protein identification tool be adopted by most researchers in proteomic studies? 4) should more tools be used to cross-compare identified proteins? and 5) is the protein sequence coverage method as implemented by most protein identification tools the only method available to interpret tandem mass spectral data? In this chapter, we showed a strategy that applies the unique peptide concept and cross-comparison method to successfully group the identified proteins into different levels of confidence.

- A species-specific two-pass algorithm to identify proteins from MS/MS spectral data was developed in our laboratory (Chapter 4). The results from the algorithm were compared to those identified by other protein identification tools, showing that our algorithm is as effective as the others.
- Protein profiles of *S. cerevisiae* grown under low and high specific gravity conditions were obtained from Mascot and our proposed two-pass approach. The confidences of identified proteins were analyzed using our developed two-step strategy and the proteins with the highest confidence were used to interpret the changes in ethanol production yield over glucose consumption under different gravity conditions (Chapter 5).

1.2.2 Other contributions

In addition to the above contributions, I have also extended the knowledge learned through this study to the following areas:

- The algorithm based on the two-pass approach was modified to take protein

phosphorylation into account and to locate the possible phosphorylation sites on phosphorylated proteins. This modified algorithm was validated using published literature data and the detailed description was published in *Proteomics* (see Appendix A).

- An automated approach to extract metabolically related proteins for metabolic flux analysis with *Pseudomonas putida* was developed. This is an example of applying bioinformatics to metabolic engineering. This work was presented at the 1st Water and Environment Specialty Conference, hosted by the Canadian Society for Civil Engineering in Saskatoon on June 2-5, 2004 (See Appendix B).

1.3 Thesis organization

This thesis consists of six chapters. Chapter 1 is a short introduction of this thesis. Chapter 2 is a literature review of the major proteomic study techniques and the applications of proteomics in biological research. Chapter 3, 4, and 5 contain the major contributions of the thesis. Chapter 3 describes the general interpretation procedures involved in several currently widely used protein identification tools and our two-step strategy developed to analyze the protein confidence or to locate the highly-confident proteins. In addition to the developed strategy, we also developed a two-pass approach to interpret tandem MS spectral data in order to identify proteins. The detailed information of this approach is provided in Chapter 4. After that, the two-pass approach and two-step strategy were used to identify and compare the protein profiles of *S. cerevisiae* grown in different stress conditions. This part was shown in the case study found in Chapter 5. For easy reading, the contents of Chapter 3, 4 and 5 are arranged in

manuscript format. Finally, the conclusions obtained from this thesis are presented in Chapter 6 along with possible directions for future work.

1.4 References

Dass, C. (2001) *Principles and Practice of Biological Mass Spectrometry*. NY: John Wiley.

Eng, J.K., McCormack, A.L. and Yates, J.R., III. (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.

Hoffmann, E.d. (2002) *Mass Spectrometry: Principles and Applications*. NY: John Wiley.

Hunter, T.C., Andon, N.L., Koller, A., Yates, J.R., III and Haynes, P.A. (2002) The functional proteomics toolbox: methods and applications. *J Chromatogr B Analyt Technol Biomed Life Sci*, **782**, 165-181.

Kinter, M. and Sherman, N.E. (2000) *Protein Sequencing and Identification using Tandem Mass Spectrometry*. NY: Wiley-Interscience.

Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551-3567.

Yarmush, M.L. and Jayaraman, A. (2002) Advances in proteomic technologies. *Annu Rev Biomed Eng*, **4**, 349-373.

Yates, J.R., III. (2004) Mass spectral analysis in proteomics. *Annu Rev Biophys Biomol Struct*, **33**, 297-316.

Yates, J.R., III, Eng, J.K., McCormack, A.L. and Schieltz, D. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem*, **67**, 1426-1436.

Zhao, Y. and Lin, Y.-H. (2003) Growth of *Saccharomyces cerevisiae* in a chemostat under high glucose conditions. *Biotechnol Lett*, **25**, 1151-1154.

Zhao, Y. and Lin, Y.-H. (2004) An automated approach to extract metabolically related proteins for metabolic flux analysis on *Pseudomonas putida*. *Proceedings of 1st Water and Environment Specialty Conference of the Canadian Society for Civil Engineering*, WE-191, 1-7.

Zhao, Y. and Lin, Y.-H. (2005a) The development of an algorithm for the mass spectral interpretation of phosphoproteins. *Proteomics*, **5,** 843-845.

Zhao, Y. and Lin, Y.-H. (2005b) A proteomic tool for the protein identification from the tandem mass spectral data. *Proteomics*, **5,** 853-855.

Chapter 2 Proteomics and applications

In this chapter, the major concepts and techniques related to proteomics are described, including: 1) two dimensional polyacrylamide gel electrophoresis (2D-PAGE); 2) mass spectrometry for protein identification; 3) multi-dimensional protein identification technology (MudPIT); 4) protein identification tools; and 5) current major applications of proteomics.

2.1 Two-dimensional polyacrylamide gel electrophoresis

Two-dimensional PAGE is an important method for proteomic study, because a large number of proteins can be resolved in a single experiment (Wu and MacCoss, 2002). Protein mixtures are separated in 2D-PAGE based on two independent chemical characteristics of proteins: isoelectric point (pI) and molecular weight (MW).

Proteins are amphoteric molecules; a protein may carry a positive, negative, or a zero net charge, depending on its surrounding pH and amino acid content. The specific pH value at which the net charge of a protein is zero is called the protein's isoelectric point. In 2D-PAGE, the first dimensional protein separation is accomplished using a gel with a pH gradient based on the proteins' specific pI values. Proteins carry positive charges when the pH value is below their pIs; in the presence of an electric field,

they migrate toward the cathode. In contrast, proteins carry negative charges when the pH value is above their pIs; they then migrate toward the anode. As proteins migrate to a specific position where the pH value equals their pIs, their charge states reach neutrality and their migrations in the gel stop; as a result, proteins with similar pIs are separated from the others. This process is also referred to as isoelectric focusing (IEF), which allows proteins to be separated and concentrated on the basis of very small charge differences.

In the second dimensional separation, the separated proteins from IEF are separated orthogonally by polyacrylamide gel electrophoresis in the presence of sodium dodecyl sulphate (SDS). In this process, proteins are separated on the basis of their MW. Since there is likely no protein that has both the same pI and MW, the protein mixtures can be further separated after 2D-PAGE. The separated proteins can then be visualized by numerous staining methods such as silver and Coomassie blue to produce a two-dimensional image array, and then thousands of proteins can be identified on the basis of specific pI values and molecular weights for stained spots (Wu and MacCoss, 2002).

Two-dimensional PAGE was first introduced by O'Farrell (1975). In the original technique, the pH gradient was formed by carrier ampholytes and first-dimension separation (IEF separation) was performed in tube gels. However, 2D-PAGE was not widely used right after its first introduction due to many limitations. Firstly, the carrier ampholytes, which were used to form the pH gradient, were mixed polymers and their characteristics were different from various suppliers, and even different between batches produced from the same manufacturer. These variations made it difficult or impossible

to reproduce IEF separation results for the same sample. Secondly, the carrier ampholyte pH gradients were unstable and tended to drift during the IEF separation. Thirdly, the tube gel containing IEF proteins had low mechanical stability; bringing challenges when transferring the IEF proteins to the SDS-PAGE slab gel. Finally, technical proficiency of identifying the stained spots in a gel also limited the application of 2D-PAGE.

These limitations have been greatly improved with the introduction of immobilized pH gradient (IPG) strips for the first dimensional separation. The IPG technique was first developed by Bjellqvist *et al*. (1982) and was pioneered into 2D-PAGE by Gorg and colleagues (Gorg *et al*., 1988a, 1988b). In an IPG strip, immobilized pH gradients are formed using two solutions; one solution contains a relatively acidic mixture of acrylamido buffers and the other solution contains a relatively basic mixture. The range of pH gradients is defined by the concentrations of the various buffers in the two solutions. The obvious advantages of IPG technology over the original carrier ampholyte-generated pH gradients are: 1) the first dimensional separation is more reproducible because the fixed pH gradient cannot drift; 2) IPG strips are easier to handle than tube gels; and 3) IPG technology can increase the pH gradient range (Bjellqvist *et al*., 1993). Furthermore, a wide variety of ready-made IPG strips with wide or narrow pH ranges are currently available from many manufacturers for reasonable prices. Using these standardized gels, it is now possible to separate protein mixtures and generate highly reproducible 2D maps. Currently, almost all 2D-PAGE projects are done exclusively with IPGs as the IEF media (Garfin, 2003). In addition to experimental development, techniques for protein identification were also improved by introducing modified staining protocols such as Coomassie blue stain (Matsui *et al*., 1999) and silver

stain (Rabilloud, 1999), computer imaging analysis programs such as MELANIE (Appel *et al*., 1997a, 1997b), standard protein databases for many cells or organisms such as *S. cerevisiae* (Goffeau *et al*., 1996) and *Escherichia coli* (Blattner *et al*., 1997), and MS techniques for subsequent analysis of proteins (Beranova-Giorgianni, 2003; Figeys *et al*., 1998; Henzel *et al*., 2003; Lahm and Langen, 2000).

While 2D-PAGE has the ability to resolve many proteins in one experiment, there are several technical drawbacks that mainly stem from the physical limitations of 2D-PAGE and visualizing techniques. These drawbacks include 1) the extreme acidic or basic proteins cannot be separated using 2D-PAGE since most 2D gels can only focus proteins with a pI range between 4 and 10; 2) smaller proteins (MW $\leq$ 15 KDa) or larger proteins (MW $\geq$ 200 KDa) also cannot be separated in 2D-PAGE; 3) low solubility proteins (e.g., membrane proteins) can not be identified using 2D-PAGE (Rabilloud, 1996); and 4) only higher abundance proteins can be observed in the 2D gel while lower abundance proteins are often not seen on the gel due to the low dynamic range of typical stain techniques (Gygi *et al*., 2000). Furthermore, the various staining protocols for 2D gel also limit subsequent analysis (e.g., MALDI-MS) of the separated proteins.

Recently, many efforts have been made to overcome the above disadvantages of 2D-PAGE. Examples include 1) using very narrow pH gradients (e.g., 1 pH unit over an 18-cm gel) for IEF separation to improve the resolution and detect low abundance proteins (Gorg *et al*., 2004); 2) choosing organic solvents to aid in solubilizing hydrophobic proteins (Molloy *et al*., 1999) for 2D-PAGE; and 3) using a fluorescent stain technique to improve the sensitivity and the linear dynamic range of detection (Patton, 2000). The

fluorescent stain technique is also more compatible with subsequent analysis such as MALDI-MS compared with the traditional staining methods such as Coomassie blue and silver staining (Lauber *et al*., 2001; Patton, 2002). In addition to those improvements for 2D-PAGE, protein enrichment approaches such as sequential extraction (Bae *et al*., 2003) and affinity chromatography separation (Lee and Lee, 2004) are also applied to enrich basic or hydrophobic proteins prior to 2D-PAGE analysis.

2.2 Mass spectrometry for protein analysis

Mass spectrometry (MS) is an instrumental approach for separating and measuring molecular ions according to their mass-to-charge ratio (m/z). MS can provide both the molecular mass and structural information of an ion of interest; it is also applicable to samples with a wide variety of characteristics (e.g., volatile, non-volatile, polar, nonpolar, and so on) (Dass, 2001). A typical MS has three basic components (Figure 2.1A): 1) an ion source that converts the neutral sample molecules into gas-phase ions; 2) a mass analyzer that separates and mass-analyzes ionic species; and 3) a detector that measures the relative abundance of the mass-resolved ions. Tandem mass spectrometry (MS/MS) involves the use of two or more mass analyzers. It is often used to analyze individual components of a mixture. A major difference between MS/MS and MS is that a collision-induced dissociation (CID, also called Q_2 in a triple quadrupole mass spectrometer) chamber is used to connect two mass analyzers (Q_1 and Q_3, also called MS-1 and MS-2, respectively) in MS/MS (Figure 2.1B). The function of the CID is to dissociate a pre-selected ion into smaller fragments by collision of the pre-selected ion with inert gas molecules (e.g., argon) with the aid of collision energy.

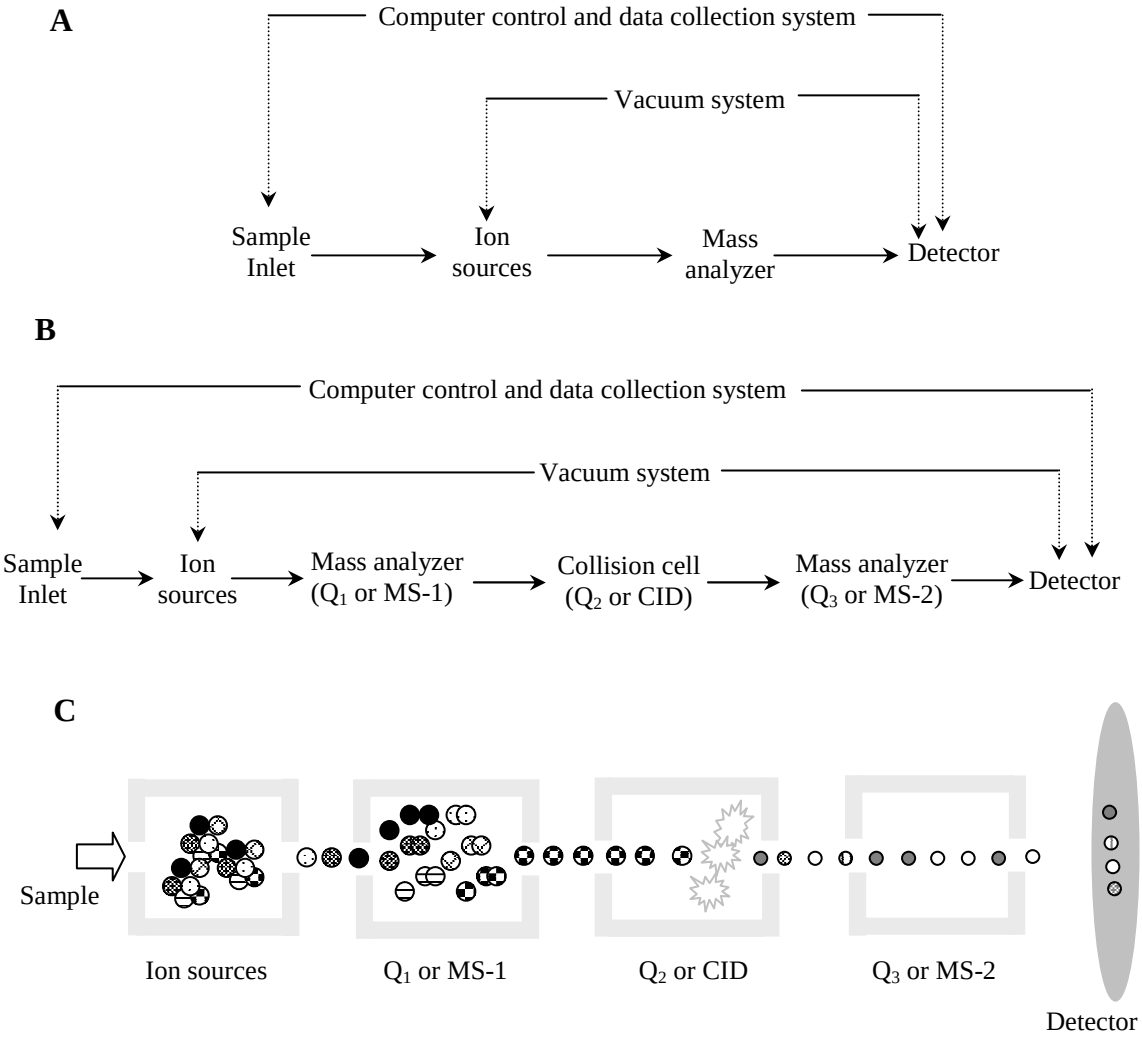


Figure 2.1 Fundamental components of a mass spectrometer
A: components of a single MS
B: components of a MS/MS (triple quadruple as an example)
C: schematic of MS/MS operation

In MS-based proteomic studies, MS operation can be grouped into two categories: single stage MS and tandem MS based operation. In single stage MS operation, the first step is to convert the analyte molecules into gas-phase ionic species and dissociated fragments; then a mass analyzer separates these molecular ions and their charged fragments according to their m/z; and finally the separated ions are detected by a detector and displayed in the form of a mass spectrum. After that, the molecular mass and structure can be derived from the information of the spectrum. In proteomic studies, single stage MS is generally used to identify the structure or sequence of a single or purified protein. For example, a MALDI-MS is used to study proteins separated by 2D-PAGE.

Tandem mass spectrometers can analyze a more complex sample such as a protein mixture. MS/MS are generally operated in four modes: product ion scan, precursor ion scan, neutral loss scan, and selected-reaction monitoring (Arnott, 2001; Dass, 2001). The most common operational mode for proteomic analysis is product ion scan. The concept of product ion scan operation is illustrated in Figure 2.1C, and involves mass-selection, fragmentation, and mass analysis. These three steps are performed using two stages of mass analysis.

Firstly, mass analysis is solely performed using Q_1, while Q_2 and Q_3 are set to only transmit ions to the detector. As a result, a mass spectrum is obtained after the Q_1 scan (so called MS data). A researcher can then select the ion of interest from the MS data and set Q_1 to transmit only the selected ion for subsequent structural determination using Q_3. By definition, the selected ion is called the precursor ion (the former term was the “parent” ion).

Secondly, the selected ion is transferred into the CID chamber (Q_2) for fragmentation via collisions with inert gas atoms. Generally, the selected ions are peptides in proteomic analysis using MS/MS. For the peptides undergoing low-energy CID, a series of fragments that contain the N-terminal or C-terminal portions of the peptide are produced. For the N-terminal fragment, the ion is classed as either a, b or c, depending on the cleaved bond. For a C-terminal fragment, the ion type is either x, y or z. A subscript indicates the number of residues in the fragment, for example a_2, b_2, and so on (Figure 2.2). The nomenclature for fragment ions was proposed by Roepstorff and Fohlman (1984). Some of the fragments may also undergo neutral losses of small molecules, such as ammonia or water to form peaks with 17 Da (ammonia) or 18 Da (water) reduction (Ballard and Gaskell, 1993). By definition, these fragmented ions are called product ions (previously called "daughter" ions). The low-energy CID is particularly useful in the analysis of peptides, because the fragmentation frequently occurs at amide bonds, so the peptide's sequence can be characterized from the product ions (Hunt *et al*., 1986).

Finally, the product ions generated from CID are scanned at Q_3 and detected by an ion detector. The collections of mass information of these product ions are called MS/MS spectral data. Since the MS/MS spectral data contain the structural information of the peptide of interest, the peptide may then be identified with the aid of computational software tools. One set of MS/MS spectral data corresponds to one precursor ion. By resetting MS data for another pre-selected precursor ion, a new set of MS/MS spectral

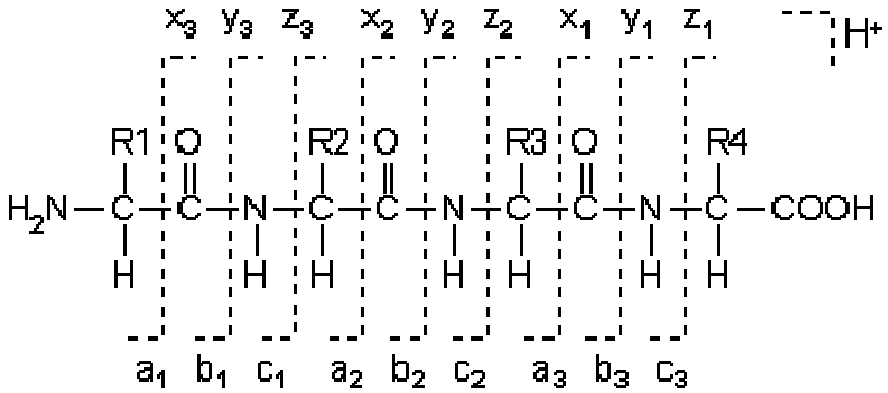


Figure 2.2 Schematic of peptide fragmentation (http://www.matrixscience.com/help/fragmentation_help.html)

data is obtained after the Q_3 scan. As a result, another peptide in the sample mixture can be identified.

This tandem mass operation is akin to the combination of a chromatography technique with a single MS. The first stage of MS/MS operation separates a mixture of ions according to the mass of individual components, in the same fashion as the chromatography technique resolves a mixture of compounds; the second stage of MS/MS operation obtains mass spectra of each mass-resolved ion as the single stage operation of MS.

The success of MS/MS for proteomic study depends on four criteria. The first criterion is mass detection range, which is defined as the maximum allowable mass that can be analyzed. The second criterion is detection sensitivity, which is defined as the smallest amount of an analyte that can be detected at a certain confidence level. In proteomic work, instruments that are routinely capable of obtaining data on femtomole (10^{-15} moles) quantities of peptides or less are recommended since the amounts of proteins are generally limited. The third criterion is mass resolution, which is defined as the ability to differentiate two neighbouring mass ions. An MS/MS capable of high-resolution is very important for protein analysis. For example, when using the first MS to transmit two precursor ions that have the same nominal mass, if the resolution is low (e.g., ±1 Da), then the MS-1 (See Figure 2.1B) will simultaneously transfer these two ions into CID and MS-2 (See Figure 2.1B), subsequently making the MS/MS spectral data difficult to interpret. In contrast, if the resolution is high, the MS-1 can separate these ions and

transfer them one by one, such that the generated MS/MS spectrum is strictly linked to one precursor ion, increasing the confidence of subsequent protein identification. The last criterion is the accuracy of mass measurement. In proteomic studies, the measured values for peptide ions or their fragments must be as close as possible to their real values. This is particularly useful when the MS or MS/MS data are subsequently used to search the peptide sequences in a reference peptide database.

2.3 Multi-dimensional protein identification technology

Multi-dimensional protein identification technology (MudPIT) is used to analyze proteins using liquid chromatography coupled with tandem mass spectrometry. Unlike the 2D-PAGE technique, in which intact proteins are separated and identified, MudPIT separates and identifies peptide mixtures digested from protein mixtures by a specific enzyme. Trypsin, for example, is the commonly used digestive enzyme and it selectively cleaves proteins by cutting at lysine and arginine residues (except those next to proline) (Kinter and Sherman, 2000; Synder, 2000), yielding a number of peptides (so called tryptic peptides) with different lengths and amino acid sequences.

MudPIT involves several steps: 1) fractioning peptide mixtures by loading the mixture onto a strong cation exchange (SCX) column and then eluting the column with salt gradients in order to separate peptides according to their charge and produce a series of peptide fractions; 2) separating the mixture of peptides in each collected fraction by loading the sample onto a reverse phase (RP) column and then eluting the column with a polar solvent (e.g., water mixed with methanol or acetonitrile) gradient to separate the peptides based on hydrophobicity; 3)

transferring the peptides separated from step 2 into a tandem mass spectrometer (MS/MS) for detection using product ion scan mode (see Section 2.2 for detail); 4) interpreting the MS/MS spectral data from all of the fractions for peptide and protein identification (Eng *et al*., 1994; Yates *et al*., 1995).

In MS/MS analysis, peptides are separated according to their specific mass (as an extra "dimension of separation"), so MudPIT includes at least three steps for peptide separation, resulting in higher resolution than 2D-PAGE. MudPIT can be accomplished in either off-line or on-line modes. In off-line operation, firstly the peptide mixtures are prefractionated using an SCX column, and then the resultant peptide fraction is separated using a reverse phase column and the isolated peptides are transferred into MS/MS for determination. Typically, the prefraction by SCX column is operated independently, while the reverse phase chromatography separation and the tandem mass spectrometric analysis are coupled together (referred to as LC-MS/MS) (Peng *et al*., 2003; Pflieger *et al*., 2002). The on-line operation integrates the SCX column prefraction, reverse phase separation, and tandem mass spectrometric analysis, (referred to as LC-LC-MS/MS) (Huang *et al*., 2000; Mawuenyega *et al*., 2003; Mitulovic *et al*., 2004; Wagner *et al*., 2002). The obvious advantage of MudPIT over 2D-PAGE is that the former greatly increases the number of identified peptides and proteins because MudPIT can detect proteins over a wide range of pI, abundance, and subcellular localization such as membrane, ribosome (Koller *et al*., 2002; Link *et al*., 1999; Washburn *et al*., 2001). The second advantage is that MudPIT can be fully automated.

The main weakness of MudPIT is with post-experimental data processing. During a MudPIT experiment, a set of MS/MS spectral data representing product ions is collected to identify one corresponding precursor ion (peptide ion in the peptide mixtures), so there will be an extremely large volume of mass spectral data waiting for interpretation after MudPIT. This is particularly true when a whole cell proteomic investigation is conducted. For example, Peng *et al*. (2003) recorded more than 162,000 mass spectra when doing yeast proteome analysis using an off-line MudPIT. The tremendous amount of mass spectral data presents a significant problem in terms of the time required to assign the collected data into a useable format for subsequent protein identification. In addition, the success of MudPIT relies on the availability of a complete sequenced genome of interested cells or organisms for protein identification. Finally, the MudPIT instrument is expensive and requires dedicated personnel. This therefore limits accessibility of the instrument by others.

Nevertheless, MudPIT is the best alternative technique to 2D-PAGE. The problems may be alleviated over time. For example, computing resources will continue to steadily increase in performance and become more affordable; the mass spectrometric instrumentation, computer algorithms for MS/MS spectral data interpretation, and genomic sequence data for major research organisms will also surely improve. These improvements may eliminate the disadvantages of MudPIT in the future, and MudPIT will clearly become an increasingly attractive tool.

2.4 Protein identification tools

Mass spectrometry (MS) and tandem mass spectrometry (MS/MS) have been the major instruments for proteomic studies. Currently, two strategies are widely used. The first one is to separate protein mixtures by 2D-PAGE. Then the protein(s) of interest is cut out and digested by a specific enzyme (e.g., trypsin) to generate a series of peptides. Then, these peptides are analyzed by MS. Since all peptides are generated from the same protein, the protein can be identified after interpreting all the corresponding MS spectral data (Henzel *et al*., 2003; Jensen *et al*., 1997; Yates *et al*., 1993). This strategy is referred to as peptide mass fingerprinting (PMF). The second strategy is to enzymatically digest proteins in the original sample before the separation step; and then separate the proteolytic peptides by LC coupled with an ion exchange column (e.g., SCX) to generate a series of peptide fraction. The resultant peptides in each fraction are then separated by an LC coupled with a reverse phase column, followed by MS/MS analysis (McCormack *et al*., 1997; Yates *et al*., 1999, 2000). During the MS/MS analysis, each precursor ion (representing a peptide) is subjected to selection, fragmentation, and sequence determination. The identified peptides from MS/MS analysis are then collected to identify proteins. Since the peptides are identified by a series of product ions generated through fragmentation in the MS/MS analysis, this strategy is referred to as peptide fragmentation fingerprinting (PFF).

No matter which strategy is used in proteomic studies, the critical step is to interpret experimentally generated MS or MS/MS spectral data for protein identification (so called post-experimental data processing). Generally speaking, both MS and MS/MS

spectra can be manually interpreted on the basis of m/z values for each peak and/or the difference between close peaks in the spectrum (Hoffmann, 2002; Staudenmann and James, 2001; Synder, 2000). A peak with m/z = 115 Da and z = 1, for example, represents aspartic acid (single letter form: D; see the list of amino acids for detail), while m/z = 186 Da may represent four choices of amino acid residues or combinations at z = 1, including tryptophan (W), glycine-glutamic acid (G-E), alanine-aspartic acid (A-D), or serine-valine (S-V). Theoretically, a protein sequence or peptide sequence can be predicted after correlating all m/z data in a spectrum to amino acids. However, some important drawbacks limit this method to practical application. Firstly, the amino acids isoleucine (I) and leucine (L) have the same m/z value (their nominal value is the same) at z = 1, making it difficult to decipher which one really exists when identifying an amino acid from an m/z of 113 Da. Secondly, an m/z may represent several combinations of amino acids like the example of m/z = 186 Da illustrated above. Thirdly, different kinds of product ions (e.g., b- and y- type) are present simultaneously in an MS/MS spectrum; and it is not easy to differentiate them. Fourthly, not all product ions can be detected by MS or MS/MS, so it is not possible to identify the 'true' protein sequence that matched the reference one found in the database. Finally, even though manual interpretation on the MS or MS/MS spectra can be successfully achieved, it requires a tremendous amount of effort. In fact, proteomic analysis would be impossible if software tools were not available to interpret the generated MS and MS/MS data for protein identification (Gygi and Aebersold, 2000).

Publicly available genome sequence information makes it possible to automatically interpret mass spectral data for protein identification by providing standard protein

sequence databases. In fact, all currently available software tools for MS and MS/MS spectra interpretation are designed on the basis of genome sequences of specific species. Generally, a species-specific protein database is obtained from a public source in any identification operation. According to the protein analysis process, the existing post-experimental data processing tools can be grouped into PMF tools and PFF tools (Table 2.1). Some of these software packages are publicly accessible and others are commercial products.

Table 2.1 Software tools for MS spectra interpretation

Peptide mass fingerprinting (PMF) searching tools		
Tools	Website	Reference
Mascot	http://www.matrixscience.com	(Perkins *et al*., 1999)
Mowse	http://www.hgmp.mrc.ac.uk	(Pappin *et al*., 1993)
MS-Fit	http://www.prospector.ucsf.edu	(Clauser *et al*., 1995)
PepSea	http://www.pepsea.protana.com	(Mann *et al*., 1993; Mann and Wilm, 1994)
ProFound	http://www.proteometrics.com	(Zhang and Chait, 2000)
PeptIdent / MultiIdent	http://www.expasy.ch/tools	(Wilkins *et al*., 1999; Wilkins *et al*., 1998)
Peptide fragmentation fingerprinting (PFF) searching tools		
Tools	Website	Reference
Mascot	http://www.matrixscience.com	(Perkins *et al*., 1999)
MS-Tag	http://www.prospector.ucsf.edu	(Clauser *et al*., 1999)
PepSea	http://www.pepsea.protana.com	(Mann and Wilm, 1994)
SEQUEST	http://www.fields.scripps.edu/sequest	(Eng *et al*., 1994; Yates *et al*., 1995)
PepFrag	http://www.proteometrics.com	(Fenyo *et al*., 1998)

In PMF analysis, each protein sequence in the database is 'computer-digested' according to the specificity of the enzyme, and the masses of the resulting peptides are calculated. Then the masses of experimentally measured proteolytic peptides (so called MS data) are compared to the theoretical masses of computer-proteolysis peptides. Generally it is requested that at least three to six matched peptides are derived from the same protein in order to positively identify a protein, even though it is reported that only a few determined peptides are sufficient for identification of a protein when the genome sequence is available (Fenyo, 2000). In the theoretical peptide database, it is common that there are several peptides from different proteins that have the same nominal m/z, representing multiple choices for an experimental peak in an MS spectrum. Thus several proteins are typically predicted from an MS spectrum as potential candidates. A score, therefore, is needed to qualify each candidate. Generally, the score is calculated during the comparison between the experimental peptides with theoretical peptides; the possible protein sequences are sorted according to the score and the protein sequence with the highest score is selected as the identified protein. The recent development in higher mass accuracy MS has improved the success rate for protein identification by PMF (Clauser *et al*., 1999). However, the application of PMF is usually limited to pure proteins or simple protein mixtures (Zhang and Chait, 2000).

In PFF analysis, the first step is to generate a database containing peptide sequences and their corresponding mass; this step is similar to the first step in PMF. The m/z of a selected precursor ion in MS/MS analysis is then used to find all possible peptides from the peptide database. After that, each peptide candidate sequence is computer-dissociated by simulating the fragmentation in CID to generate a theoretical fragment

mass spectrum. Then the theoretical spectrum is compared to the measured fragment mass spectrum (MS/MS data). Like the PMF, a score qualifying the comparison is calculated and used to sort the peptide candidates; the peptide with the highest score is normally considered as the identified result (Eng *et al*., 1994; Mann and Wilm, 1994). In contrast to PMF, PFF analysis provides the amino acid sequence of each peptide, and this information enables the identification of a protein from a single peptide (e.g., identified peptides consisting of more than ten amino acids, Eng *et al*., 1994). PFF analysis has proved to be more useful in protein identification than PMF analysis (Yarmush and Jayaraman, 2002), and it is also the best choice for identifying complex protein mixtures.

During the application of protein identification tools, the first step is to predict the protein or peptide from the experimental PMF or PFF data by searching against a specific database. Some searching parameters are provided on the basis of experimental conditions. Generally, these parameters include the choice of searched ions (e.g., monoisotopic ion or average ion), mass tolerance, charge state of precursor ions, cysteine modifications, and the ranges of pI and MW of protein candidates. The second step is to analyze the searched results and rank them. The key problem in mass spectrometry-based protein identification is that each measured mass can randomly match a series of peptides from a specific sequence database. This is to say the peptides determined by protein identification tools for the same mass spectral data are generally not unique. Therefore, software tools for PMF and PFF protein identification must implement scoring strategies to distinguish the most probable peptide (protein) from the others.

The simplest and most obvious scoring method is to count the number of matched peptides (in PMF) or product ions (in PFF) between measured peptide or product ion and theoretically calculated peptide or product. The searched proteins or peptides are then ranked according to the matching number. The software tools applying this method include PepSea, PeptIdent/MultiIdent, and MS-Fit for PMF analysis, PepFrag and MS-Tag for PFF analysis.

The sophisticated methods for identifying proteins are based on statistical analysis. For example, Mowse takes into account the relative distribution frequency of peptides in the source database when calculating the score. Mascot, which implemented the Mowse, calculates the probability for each peptide sequence in the database on the basis of the observed match between experimental data and a protein sequence and the absolute probability of the protein by adding all the probabilities of its peptides. ProFound uses Bayesian theory to rank protein sequences in the database according to their probability of occurrence. SEQUEST ranks its results according to cross-correlation scores (X_{corr}), which is calculated by comparing the measured fragment mass spectra with the protein sequences in the database.

2.5 Applications of proteomics

Proteomics is an impressive and important approach in biological research. Currently, proteomic studies can be applied in four basic areas: 1) protein profiling; 2) protein quantification; 3) mapping of modified proteins; and 4) mapping of a protein-network.

2.5.1 Protein profiling

2.5.1.1 Protein mining

One goal of protein profiling is to identify all (or as many as possible) of the proteins expressed in a cell or tissue sample. This process is also called protein mining. Because of the potential for high throughput analysis, mass spectrometry is now routinely used to identify proteins (e.g., PMF analysis) separated by 2D-PAGE (Gorg *et al.*, 2004). Along with a database search, the 2D-PAGE method has been used to construct proteome maps for many organisms such as *Escherichia coli* (Tonella *et al*., 1998), *Shigella flexneri* (Liao *et al*., 2003), *Salmonella enteritidis* (Park *et al*., 2003), *Saccharomyces cerevisiae* (Perrot *et al*., 1999), *Caenorhabditis elegans* (Schrimpf *et al*., 2001), and various other samples such as lymphoblastoid B-cell (Caron *et al*., 2002), human macrophage (Dupont *et al*., 2004), and murine R1 embryonic stem cells (Elliott *et al*., 2004). These proteome maps serve as databases for further comparative proteomic analysis (Cordwell *et al*., 1999; Pleissner *et al*., 2004).

2.5.1.2 Comparative proteomic analysis

Another goal of protein profiling is to study cellular responses and adaptation mechanisms when a cell is exposed to various conditions. This task can be accomplished using comparative proteomic analysis. Comparative proteomic analysis is a method to study a cell or organism grown at a particular state (e.g., various growth state or disease state) or subjected to a particular stimulus (e.g., nutrients, chemicals, or drugs) by

comparing the expressed protein profiling of this cell or organism with that of a normal cell or organism. The proteins expressed differently in the two samples are very important for interpreting some observed behaviour phenomenon (so called phenotype) of the abnormal cell or organism. The comparative proteomic analysis is also referred to as differential display proteomics (Cordwell *et al*., 2001).

Two-dimensional PAGE is particularly well suited for comparative proteomic analysis, because it not only resolves many proteins reproducibly, but also provides intact proteins of interest for subsequent analysis. After 2D-PAGE separation and staining, image software is used to analyze the color spots on the gel page. The different proteins between two samples can be detected and identified by their pIs and MWs by searching against available databases. If some proteins cannot be identified, the spots can be cut out and subjected to subsequent analysis (determination of the sequence, structure, and function of these proteins). Examples of studies comparing proteins of a cell include: different nutrient conditions (Franzen *et al*., 1999), heat stress conditions (Periago *et al*., 2002a, 2002b), and drug treatment responses (Fountoulakis *et al*., 2000).

2.5.2 Protein quantification

An increasing emphasis in proteomics is the quantification of protein content rather than simple determination of presence or absence (Wu and MacCoss, 2002). Quantification of protein(s) is carried out by analyzing the level of change of the interested protein(s) expressed by a cell grown at a particular state compared to those protein(s) expressed by a cell under “normal” conditions. This is actually a specialized form of comparative

proteomic analysis, because the essential condition for protein quantification is that the targeted proteins must be expressed by both samples. Technically, it is difficult to measure the absolute quantities of peptides, because peptide standards are not suitable for all kinds of peptides in large-scale measurements. Therefore, the current proteome-wide protein quantification is still a comparative study (the detection of up- or down-regulated proteins). This can be successfully done using both 2D-PAGE and MudPIT approaches.

After 2D-PAGE separation and staining, the protein levels in two samples can be quantitatively analyzed by comparing the intensities of stained spots in 2D-PAGE gels. The affected proteins at the particular condition can be assessed simultaneously; this may be one reason that the 2D-PAGE technique is still an important tool in proteome analysis (Rabilloud, 2002). However, the major technical limitation is that it is difficult to exactly match protein spots between two independent 2D-PAGE images due to the inherent variability in 2D-PAGE separation. For example, streaking of the spots and/or bending in the gel can result in variability between 2D-PAGE gels, making it a difficult, laborious task to compare gel images (Blomberg *et al*., 1995). A direct approach to overcome this inherent inter-gel variability is to separate and compare two protein samples in the same gel. The proteins from different conditions are variously labelled, then the separated proteins are analyzed by advanced software tools and the difference between the same proteins under different conditions can be measured. This method is referred to as differential gel electrophoresis (Knowles *et al*., 2003; Monribot-Espagne and Boucherie, 2002). Following are several examples to illustrate how this process works.

The first example is to use two different fluorescent dyes to label *in vitro* the two protein samples prior to the first dimension 2D-PAGE. Then, these samples are combined and separated using the same first and second dimension gels, and finally the gel images are visualized using fluorescent scanning at the two separate wavelengths specific to the two fluorescent dyes (Tonge *et al*., 2001; Yan *et al*., 2002). This technique enables the proteins present in each of the original samples to be viewed separately and makes even subtle differences in protein expression levels immediately apparent. The second example is to radio-label one sample with ^{14}C and the other one with ^{3}H, and then combines the samples and separates them by 2D-PAGE in the same gel. Finally the $^{3}H/^{14}C$ ratio of each protein spot is determined by exposure to two types of imaging plates, one sensitive to ^{14}C and the other to both ^{14}C and ^{3}H (Monribot-Espagne and Boucherie, 2002). The last example is using stable-isotopes to label samples followed by quantification analysis by MS/MS, as MS/MS has the ability to differentiate the change in mass of a protein or peptide that is introduced by a stable-isotope during cell culture. Firstly, one cell sample is grown on medium containing the naturally occurring abundance of stable-isotopes ^{14}N (99.6%) and ^{15}N (0.4%), while a second sample is grown on the same medium enriched in ^{15}N (>96%). Then the two sample pools are combined and the resulting proteins are separated using 2D-PAGE, proteolyzed using trypsin, and analyzed using MS/MS (Oda *et al*., 1999). Finally the resulting spectra are used to both identify the protein and determine the relative abundance in the two cellular protein extracts.

Protein quantification analysis can also be achieved using MudPIT. The widely used technique is called isotope-coded affinity tag (ICAT) peptide labelling (Gygi *et al*., 1999). The method consists of four major steps. Firstly, one protein sample is labelled with a light version of ICAT reagent, while the other sample is labelled with a heavy version of ICAT reagent. Both labelled samples are then combined and digested. The tagged peptides are then isolated by avidin affinity chromatography, and the isolated tagged peptides are separated and analyzed by capillary LC-MS/MS.

2.5.3 Mapping of protein modification

An important application of proteomics is the characterization of protein(s) with post-translational modification (PTM). PTM proteins refer to proteins subjected to covalent modification of side chains after they are translated. The goal of PTM is to influence protein structure, target, function, and interactions with other proteins. For example, phosphorylation is found on threonine, serine and tyrosine residues, and plays a central role in the regulation of many cellular processes such as cell cycle, growth, apoptosis and differentiation. The corresponding protein is called a phosphorylated protein. There are many other kinds of modifications (e.g., acetylation, glycosylation, methylation, and so on) in a cell system (Aebersold and Mann, 2003; Mann and Jensen, 2003). The task of mapping modified protein(s) consists of identifying which protein(s) is modified, determining what kind of PTM it is, and locating which amino acid(s) are modified. Such information is important to identify cellular response mechanisms in a wide variety of biological processes and disease states such as cancer. Thus, a better understanding of these modifications would help investigators design better treatment strategies.

Traditional strategies for PTM protein mapping involve purification of protein samples and identification of any modification on each purified protein. However, this method is not suitable for proteomic studies in which a large number of proteins are to be systematically analyzed for PTM mapping. The introduction of MS methods to analyze peptides now offers a better means to characterize protein modifications, because MS methods measure both native peptide masses and their counterpart modified peptide masses to provide direct analytical data. For example, analysis of a peptide and its phosphovariant by MALDI-MS yields two signals: one at lower m/z is for the native peptide, and the other at 80 Da higher (m/z) is the corresponding phosphorylated peptide. Thus, a single MS analysis can identify the proteins and their modified forms. To predict the special site of modification, however, tandem MS must be applied. After the fragmentation of the phosphopeptide and the measurement of the masses of the resulting fragments (see Section 2.2 for the detail information of MS operation), the specific information regarding the sequence data and sites of modification can be obtained. The obvious advantages of MS methods over traditional strategy include high throughput, high sensitivity, and the ability to simultaneously analyze large number of proteins. Coupled with a powerful separation technique such as 2D-PAGE or LC-LC, MS has now become the major instrument for mapping PTM proteins.

Using 2D-PAGE as mentioned earlier, protein mixtures are separated by pI and MW, and then the modified proteins are specifically visualized on gels or on membranes. For example, phosphorylated proteins can be recognized by anti-phosphoric acid antibodies. These spots can then be excised and identified by MS and MS/MS (Aulak *et al*., 2004),

Alternatively, protein populations can be run on 2D-PAGE before and after enzymatic removal of the modifying group (e.g., alkaline phosphatase for dephosphorylation). The "disappearing" protein spots are an indication of the modification in question (Yamagata *et al.*, 2002). Similarly, the located modified proteins can be further determined using MS and MS/MS. However, the inherent drawback (e.g., inability to separate proteins with extreme pI or MW) of 2D-PAGE as introduced in Section 2.1 still limits the wide application of 2D-PAGE.

In practice, however, two factors are critical to successfully utilize MS approaches to map protein modification. The first one is the sequence coverage, which is referred to as the number of amino acids that can be identified from MS spectral data. For example, Protein P00549 is composed of 500 amino acids (Figure 2.3), which can be digested into 113 peptides by trypsin. If peptide 4 ‘LERLTSLNVVAGSDLR’ (16 amino acids) was identified, then the protein sequence coverage is 3.2%.

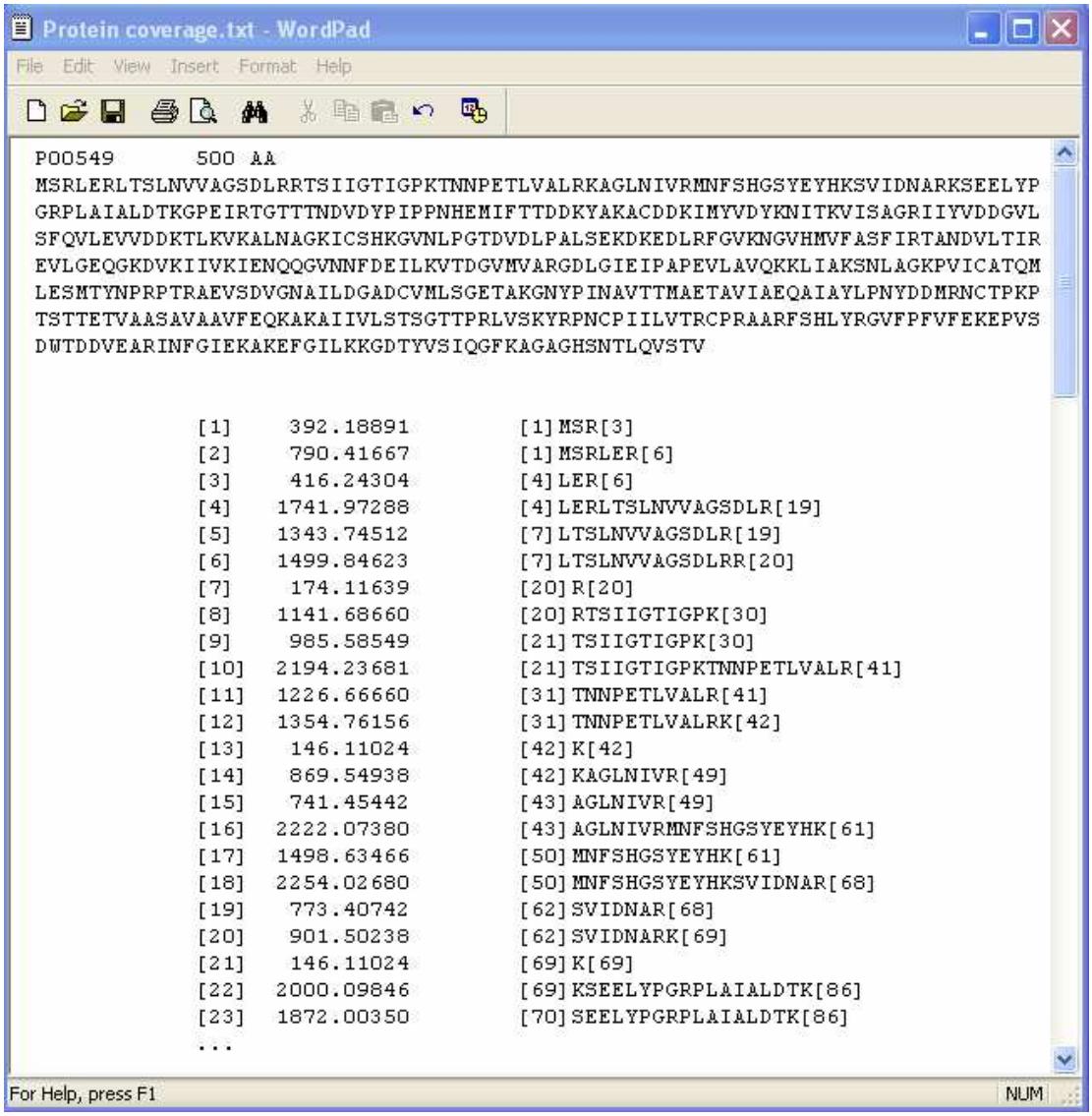


Figure 2.3 Partial listing of Protein P00549 and its peptides

In practice, there are no concrete rules to decide what the best value of protein coverage is for protein identification. Generally speaking, for the purposes of protein identification based on MALDI-MS spectral data, as little as 10–15% sequence coverage is sufficient (LoPachin *et al*., 2003). However, the coverage requirements for mapping peptide modifications by MS are far more demanding. Since many proteins contain multiple sites that could be modified, thus we need as much of the MS data as possible to identify protein modifications. That is to say 100% sequence coverage may be necessary to ensure analysis of all possible modified sites. However, this situation is very difficult or impossible to achieve when one tries to map protein modifications by MS. For example, small peptides of only a few amino acids or large peptides of more than 30 amino acids are often not detected due to a mass scan range limitation. Currently, one solution to overcome this hurdle is to do a second experiment in which proteins are digested by another enzyme with a different specificity (Mann and Jensen, 2003).

The second problem is obtaining high quality MS spectra of modified peptides. Of all the copies of any particular protein in a cell, only a small fraction, lower than 10% in many cases, may bear any specific modification (James, 2001). Therefore, strategies must be used to enrich the modified proteins or peptides in the original sample and thereby increase the probability of detection by MS. The commonly used enrichment strategies are designed based on chemical, physical or immunological properties of the modified residue. For example, the immobilized metal affinity chromatography (IMAC) method has been used for isolating phosphopeptides from protein digests since phosphopeptides can be captured selectively through their negatively charged

phosphogroup (Ficarro *et al*., 2002; Nuhse *et al*., 2003). Alternatively, antibodies directed against the modifying moiety can be used for immunoprecipitation or for immobilized antibody column chromatography. For example, proteins that had just been tyrosine-phosphorylated were immunoprecipitated with anti-phosphotyrosine antibody (Gronborg *et al*., 2002; Uljon *et al*., 2000).

2.5.4 Protein-network mapping

Most proteins carry out their functions (e.g., signal transduction, anabolism, and catabolism) in close association with other proteins by forming specific complex. Protein-network mapping is the proteomic approach to determine how proteins interact with each other in living cells. The information obtained is essential to understand how the cell functions and the consequent phenotype exerted. Basically, the protein-protein interaction can be studied using either yeast two-hybrid systems or MS approach coupled with affinity separation (Causier, 2004; Drewes and Bouwmeester, 2003).

The yeast two-hybrid approach uses a reporter gene to detect the interaction of protein pairs within the yeast cell nucleus (Fields and Sternglanz, 1994; LoPachin *et al*., 2003; Osman, 2004). Simply speaking, a protein of interest (so called bait) and a protein that might interact with the bait (so called prey) are attached to different parts of the same transcription factor. The bait protein is attached to the binding domain, whereas the prey protein is attached to the activation domain. If the proteins interact, the bait protein captures the prey protein, resulting in the re-constitution of the attached portions of the transcription factor to make it function. Subsequently a reporter gene is switched on.

Currently, the yeast two-hybrid system has been applied to mapping the protein-protein network of *E. coli* (Bartel *et al*., 1996) and *S. cerevisiae* (Uetz *et al*., 2000). However, it is important to note that the two-hybrid approaches are indirect indices of protein–protein interactions, and as a result, there are some interpretational limitations (Drewes and Bouwmeester, 2003; LoPachin *et al*., 2003).

Mass spectrometric-based proteomic analysis offers a new way to identify components of multiprotein complexes (Aebersold and Mann, 2003; Figeys *et al*., 2001; Kriwacki and Siuzdak, 2000). In this process, the associated multiprotein complexes are firstly isolated from the original protein mixture, and then the individual component of the obtained protein complex can be identified either by a 2D-PAGE separation plus MS and MS/MS analysis or by MudPIT as introduced ahead. The protein complex separation is the critical step in this process.

Two approaches are currently used to isolate multiprotein complexes. In one method, cell extracts are incubated under mild conditions with an antibody directed against one protein, the target and its interacting proteins form a protein complex and is then 'pulled down' and separated with protein mixtures (Schulze and Mann, 2004). This method is referred to as immunoprecipitation. The potential problem of this approach is the specificity of the selected antibodies. An alternative method is protein affinity chromatography, in which targeted proteins are fused to a standard affinity-tag by a generic approach as bait, which are in turn captured by antibodies or affinity resins (Rigaut *et al*., 1999). After other proteins with no specificity are washed away, the protein complex is eluted and analyzed by MS approach. This method was then further

modified to form a tandem affinity purification (TAP) technique, which utilizes two affinity tags (protein A immunoglobulin binding domains and a calmodulin binding peptide tag) to purify protein complexes that contain the TAP-tagged protein in two consecutive steps (Lee and Lee, 2004). This TAP separation procedure can give a higher yield of the purified protein complex and is now becoming the widely accepted techniques for protein-network study (Gavin *et al*., 2002; Gould *et al*., 2004; Graumann *et al*., 2004; Ho *et al*., 2002; Shevchenko *et al*., 2002).

2.6 References

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

Appel, R.D., Palagi, P.M., Walther, D., Vargas, J.R., Sanchez, J.C., Ravier, F., Pasquali, C. and Hochstrasser, D.F. (1997a) Melanie II--a third-generation software package for analysis of two-dimensional electrophoresis images: I. Features and user interface. *Electrophoresis*, **18**, 2724-2734.

Appel, R.D., Vargas, J.R., Palagi, P.M., Walther, D. and Hochstrasser, D.F. (1997b) Melanie II--a third-generation software package for analysis of two-dimensional electrophoresis images: II. Algorithms. *Electrophoresis*, **18**, 2735-2748.

Arnott, D. (2001) Basics of triple-stage quadrupole/ion-trap mass spectrometry: precursor, product and neutral-loss scanning; electrospray ionisation and nanospray ionisation. In James, P. (ed.), *Proteome Research: Mass Spectrometry*. NY: Springer.

Aulak, K.S., Koeck, T., Crabb, J.W. and Stuehr, D.J. (2004) Proteomic method for identification of tyrosine-nitrated proteins. *Methods Mol Biol*, **279**, 151-165.

Bae, S.H., Harris, A.G., Hains, P.G., Chen, H., Garfin, D.E., Hazell, S.L., Paik, Y.K., Walsh, B.J. and Cordwell, S.J. (2003) Strategies for the enrichment and identification of basic proteins in proteome projects. *Proteomics*, **3**, 569-579.

Ballard, K.D. and Gaskell, S.J. (1993) Dehydration of peptide $(M+H)^+$ ions in the gas-phase. *J Am Soc Mass Spectrom*, **4**, 477-481.

Bartel, P.L., Roecklein, J.A., SenGupta, D. and Fields, S. (1996) A protein linkage map of *Escherichia coli* bacteriophage T7. *Nat Genet*, **12**, 72-77.

Beranova-Giorgianni, S. (2003) Proteome analysis by two-dimensional gel electrophoresis and mass spectrometry: strengths and limitations. *TrAC, Trends in Analytical Chemistry*, **22**, 273-281.

Bjellqvist, B., Ek, K., Righetti, P.G., Gianazza, E., Gorg, A., Westermeier, R. and Postel, W. (1982) Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J Biochem Biophys Methods*, **6**, 317-339.

Bjellqvist, B., Pasquali, C., Ravier, F., Sanchez, J.C. and Hochstrasser, D. (1993) A nonlinear wide-range immobilized pH gradient for two-dimensional electrophoresis and its definition in a relevant pH scale. *Electrophoresis*, **14**, 1357-1365.

Blattner, F.R., Plunkett, G., III, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453-1474.

Blomberg, A., Blomberg, L., Norbeck, J., Fey, S.J., Larsen, P.M., Larsen, M., Roepstorff, P., Degand, H., Boutry, M., Posch, A. and Gorg, A. (1995) Interlaboratory reproducibility of yeast protein patterns analyzed by immobilized pH gradient two-dimensional gel electrophoresis. *Electrophoresis*, **16**, 1935-1945.

Caron, M., Imam-Sghiouar, N., Poirier, F., Le Caer, J.P., Labas, V. and Joubert-Caron, R. (2002) Proteomic map and database of lymphoblastoid proteins. *J Chromatogr B Analyt Technol Biomed Life Sci*, **771**, 197-209.

Causier, B. (2004) Studying the interactome with the yeast two-hybrid system and mass spectrometry. *Mass Spectrom Rev*, **23**, 350-367.

Clauser, K.R., Baker, P. and Burlingame, A.L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem*, **71**, 2871-2882.

Clauser, K.R., Hall, S.C., Smith, D.M., Webb, J.W., Andrews, L.E., Tran, H.M., Epstein, L.B. and Burlingame, A.L. (1995) Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melanoma proteins isolated by two-dimensional PAGE. *Proc Natl Acad Sci U S A*, **92**, 5072-5076.

Cordwell, S.J., Nouwens, A.S., Verrills, N.M., McPherson, J.C., Hains, P.G., Van Dyk, D.D. and Walsh, B.J. (1999) The microbial proteome database - an automated laboratory catalogue for monitoring protein expression in bacteria. *Electrophoresis*, **20**, 3580-3588.

Cordwell, S.J., Nouwens, A.S. and Walsh, B.J. (2001) Comparative proteomics of bacterial pathogens. *Proteomics*, **1**, 461-472.

Dass, C. (2001) *Principles and Practice of Biological Mass Spectrometry*. NY: John Wiley.

Drewes, G. and Bouwmeester, T. (2003) Global approaches to protein-protein interactions. *Curr Opin Cell Biol*, **15**, 199-205.

Dupont, A., Tokarski, C., Dekeyzer, O., Guihot, A.L., Amouyel, P., Rolando, C. and Pinet, F. (2004) Two-dimensional maps and databases of the human macrophage proteome and secretome. *Proteomics*, **4**, 1761-1778.

Elliott, S.T., Crider, D.G., Garham, C.P., Boheler, K.R. and Van Eyk, J.E. (2004) Two-dimensional gel electrophoresis database of murine R1 embryonic stem cells. *Proteomics*, ***4,*** *3813-3832*.

Eng, J.K., McCormack, A.L. and Yates, J.R., III. (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.

Fenyo, D. (2000) Identifying the proteome: software tools. *Curr Opin Biotechnol*, **11**, 391-395.

Fenyo, D., Qin, J. and Chait, B.T. (1998) Protein identification using mass spectrometric information. *Electrophoresis*, **19**, 998-1005.

Ficarro, S.B., McCleland, M.L., Stukenberg, P.T., Burke, D.J., Ross, M.M., Shabanowitz, J., Hunt, D.F. and White, F.M. (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat Biotechnol*, **20**, 301-305.

Fields, S. and Sternglanz, R. (1994) The two-hybrid system: an assay for protein-protein interactions. *Trends Genet*, **10**, 286-292.

Figeys, D., Gygi, S.P., Zhang, Y., Watts, J., Gu, M. and Aebersold, R. (1998) Electrophoresis combined with novel mass spectrometry techniques: powerful tools for the analysis of proteins and proteomes. *Electrophoresis*, **19**, 1811-1818.

Figeys, D., McBroom, L.D. and Moran, M.F. (2001) Mass spectrometry for the study of protein-protein interactions. *Methods*, **24**, 230-239.

Fountoulakis, M., Berndt, P., Boelsterli, U.A., Crameri, F., Winter, M., Albertini, S. and Suter, L. (2000) Two-dimensional database of mouse liver proteins: changes in hepatic protein levels following treatment with acetaminophen or its nontoxic regioisomer 3-acetamidophenol. *Electrophoresis*, **21**, 2148-2161.

Franzen, B., Becker, S., Mikkola, R., Tidblad, K., Tjernberg, A. and Birnbaum, S. (1999) Characterization of periplasmic *Escherichia coli* protein expression at high cell densities. *Electrophoresis*, **20**, 790-797.

Garfin, D.E. (2003) Two-dimensional gel electrophoresis: an overview. *TrAC, Trends in Analytical Chemistry*, **22**, 263-272.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141-147.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996) Life with 6000 genes. *Science*, **274**, 546, 563-547.

Gorg, A., Postel, W., Domscheit, A. and Gunther, S. (1988a) Two-dimensional electrophoresis with immobilized pH gradients of leaf proteins from barley (Hordeum vulgare): method, reproducibility and genetic aspects. *Electrophoresis*, **9**, 681-692.

Gorg, A., Postel, W. and Gunther, S. (1988b) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **9**, 531-546.

Gorg, A., Weiss, W. and Dunn, M.J. (2004) Current two-dimensional electrophoresis technology for proteomics. *Proteomics*, **4**, 3665-3685.

Gould, K.L., Ren, L., Feoktistova, A.S., Jennings, J.L. and Link, A.J. (2004) Tandem affinity purification and identification of protein complex components. *Methods*, **33**, 239-244.

Graumann, J., Dunipace, L.A., Seol, J.H., McDonald, W.H., Yates, J.R., III, Wold, B.J. and Deshaies, R.J. (2004) Applicability of tandem affinity purification MudPIT to pathway proteomics in yeast. *Mol Cell Proteomics*, **3**, 226-237.

Gronborg, M., Kristiansen, T.Z., Stensballe, A., Andersen, J.S., Ohara, O., Mann, M., Jensen, O.N. and Pandey, A. (2002) A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies: identification of a novel protein, Frigg, as a protein kinase A substrate. *Mol Cell Proteomics*, **1**, 517-527.

Gygi, S.P. and Aebersold, R. (2000) Mass spectrometry and proteomics. *Curr Opin Chem Biol*, **4**, 489-494.

Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y. and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A*, **97**, 9390-9395.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*, **17**, 994-999.

Henzel, W.J., Watanabe, C. and Stults, J.T. (2003) Protein identification: the origins of peptide mass fingerprinting. *J Am Soc Mass Spectrom*, **14**, 931-942.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D. and Tyers, M. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180-183.

Hoffmann, E.d. (2002) *Mass Spectrometry: Principles and Applications*. NY, John Wiley .

Huang, P.Q., Wall, D.B., Parus, S. and Lubman, D.M. (2000) On-line capillary liquid chromatography tandem mass spectrometry on an ion trap/reflectron time-of-flight mass spectrometer using the sequence tag database search approach for peptide sequencing and protein identification. *J Am Soc Mass Spectrom*, **11**, 127-135.

Hunt, D.F., Yates, J.R., III, Shabanowitz, J., Winston, S. and Hauer, C.R. (1986) Protein sequencing by tandem mass-spectrometry. *Proc Natl Acad Sci U S A*, **83**, 6233-6237.

James, P. (2001) Mass spectrometry and the proteome. In James, P. (ed.), *Proteome Research: Mass Spectrometry*. NY: Springer.

Jensen, O.N., Podtelejnikov, A.V. and Mann, M. (1997) Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal Chem*, **69**, 4741-4750.

Kinter, M. and Sherman, N.E. (2000) *Protein Sequencing and Identification Using Tandem Mass Spectrometry*. NY: Wiley-Interscience.

Knowles, M.R., Cervino, S., Skynner, H.A., Hunt, S.P., de Felipe, C., Salim, K., Meneses-Lorente, G., McAllister, G. and Guest, P.C. (2003) Multiplex proteomic analysis by two-dimensional differential in-gel electrophoresis. *Proteomics*, **3**, 1162-1171.

Koller, A., Washburn, M.P., Lange, B.M., Andon, N.L., Deciu, C., Haynes, P.A., Hays, L., Schieltz, D., Ulaszek, R., Wei, J., Wolters, D. and Yates, J.R., III. (2002) Proteomic survey of metabolic pathways in rice. *Proc Natl Acad Sci U S A*, **99**, 11969-11974.

Kriwacki, R.W. and Siuzdak, G. (2000) Probing protein - protein interactions with mass spectrometry. In Chapman, J.R. (ed.), *Mass Spectrometry of Proteins and Peptides*. NJ: Humana Press.

Lahm, H.W. and Langen, H. (2000) Mass spectrometry: a tool for the identification of proteins separated by gels. *Electrophoresis*, **21**, 2105-2114.

Lauber, W.M., Carroll, J.A., Dufield, D.R., Kiesel, J.R., Radabaugh, M.R. and Malone, J.P. (2001) Mass spectrometry compatibility of two-dimensional gel protein stains. *Electrophoresis*, **22**, 906-918.

Lee, W.C. and Lee, K.H. (2004) Applications of affinity chromatography in proteomics. *Anal Biochem*, **324**, 1-10.

Liao, X., Ying, T., Wang, H., Wang, J., Shi, Z., Feng, E., Wei, K., Wang, Y., Zhang, X., Huang, L., Su, G. and Huang, P. (2003) A two-dimensional proteome map of *Shigella flexneri*. *Electrophoresis*, **24**, 2864-2882.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M. and Yates, J.R., III. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol*, **17**, 676-682.

LoPachin, R.M., Jones, R.C., Patterson, T.A., Slikker, W., Jr. and Barber, D.S. (2003) Application of proteomics to the study of molecular mechanisms in neurotoxicology. *Neurotoxicology*, **24**, 761-775.

Mann, M., Hojrup, P. and Roepstorff, P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom*, **22**, 338-345.

Mann, M. and Jensen, O.N. (2003) Proteomic analysis of post-translational modifications. *Nat Biotechnol*, **21**, 255-261.

Mann, M. and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem*, **66**, 4390-4399.

Matsui, N.M., Smith-Beckerman, D.M. and Epstein, L.B. (1999) Staining of preparative 2-D gels. Coomassie blue and imidazole-zinc negative staining. *Methods Mol Biol*, **112**, 307-311.

Mawuenyega, K.G., Kaji, H., Yamuchi, Y., Shinkawa, T., Saito, H., Taoka, M., Takahashi, N. and Isobe, T. (2003) Large-scale identification of *Caenorhabditis elegans* proteins by multidimensional liquid chromatography-tandem mass spectrometry. *J Proteome Res*, **2**, 23-35.

McCormack, A.L., Schieltz, D.M., Goode, B., Yang, S., Barnes, G., Drubin, D. and Yates, J.R., III. (1997) Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal Chem*, **69**, 767-776.

Mitulovic, G., Stingl, C., Smoluch, M., Swart, R., Chervet, J.P., Steinmacher, I., Gerner, C. and Mechtler, K. (2004) Automated, on-line two-dimensional nano liquid chromatography tandem mass spectrometry for rapid analysis of complex protein digests. *Proteomics*, **4**, 2545-2557.

Molloy, M.P., Herbert, B.R., Williams, K.L. and Gooley, A.A. (1999) Extraction of *Escherichia coli* proteins with organic solvents prior to two-dimensional electrophoresis. *Electrophoresis*, **20**, 701-704.

Monribot-Espagne, C. and Boucherie, H. (2002) Differential gel exposure, a new methodology for the two-dimensional comparison of protein samples. *Proteomics*, **2**, 229-240.

Nuhse, T.S., Stensballe, A., Jensen, O.N. and Peck, S.C. (2003) Large-scale analysis of *in vivo* phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. *Mol Cell Proteomics*, **2**, 1234-1243.

Oda, Y., Huang, K., Cross, F.R., Cowburn, D. and Chait, B.T. (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A*, **96**, 6591-6596.

O'Farrell, P.H. (1975) High resolution two-dimensional electrophoresis of proteins. *J Biol Chem*, **250**, 4007-4021.

Osman, A. (2004) Yeast two-hybrid assay for studying protein-protein interactions. *Methods Mol Biol*, **270**, 403-422.

Pappin, D.J., Hojrup, P. and Bleasby, A.J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol*, **3**, 327-332.

Park, M.R., Lee, E.G., Kim, Y.H., Jung, T.S., Shin, Y.S., Shin, G.W., Cha, H.G. and Kim, G.S. (2003) Reference map of soluble proteins from *Salmonella enterica* serovar Enteritidis by two-dimensional electrophoresis. *J Vet Sci*, **4**, 143-149.

Patton, W.F. (2000) A thousand points of light: the application of fluorescence detection technologies to two-dimensional gel electrophoresis and proteomics. *Electrophoresis*, **21**, 1123-1144.

Patton, W.F. (2002) Detection technologies in proteome analysis. *J Chromatogr B Analyt Technol Biomed Life Sci*, **771**, 3-31.

Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. and 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.

Periago, P.M., Abee, T. and Wouters, J.A. (2002a) Analysis of the heat-adaptive response of psychrotrophic *Bacillus weihenstephanensis*. *Int J Food Microbiol*, **79**, 17-26.

Periago, P.M., van Schaik, W., Abee, T. and Wouters, J.A. (2002b) Identification of proteins involved in the heat stress response of *Bacillus cereus* ATCC 14579. *Appl Environ Microbiol*, **68**, 3486-3495.

Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551-3567.

Perrot, M., Sagliocco, F., Mini, T., Monribot, C., Schneider, U., Shevchenko, A., Mann, M., Jeno, P. and Boucherie, H. (1999) Two-dimensional gel protein database of *Saccharomyces cerevisiae* (update 1999). *Electrophoresis*, **20**, 2280-2298.

Pflieger, D., Le Caer, J.P., Lemaire, C., Bernard, B.A., Dujardin, G. and Rossier, J. (2002) Systematic identification of mitochondrial proteins by LC-MS/MS. *Anal Chem*, **74**, 2400-2406.

Pleissner, K.P., Eifert, T., Buettner, S., Schmidt, F., Boehme, M., Meyer, T.F., Kaufmann, S.H. and Jungblut, P.R. (2004) Web-accessible proteome databases for microbial research. *Proteomics*, **4**, 1305-1313.

Rabilloud, T. (1996) Solubilization of proteins for electrophoretic analyses. *Electrophoresis*, **17**, 813-829.

Rabilloud, T. (1999) Silver staining of 2-D electrophoresis gels. *Methods Mol Biol*, **112**, 297-305.

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

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. and Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, **17**, 1030-1032.

Roepstorff, P. and Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom*, **11**, 601.

Schrimpf, S.P., Langen, H., Gomes, A.V. and Wahlestedt, C. (2001) A two-dimensional protein map of *Caenorhabditis elegans*. *Electrophoresis*, **22**, 1224-1232.

Schulze, W.X. and Mann, M. (2004) A novel proteomic screen for peptide-protein interactions. *J Biol Chem*, **279**, 10756-10764.

Shevchenko, A., Schaft, D., Roguev, A., Pijnappel, W.W. and Stewart, A.F. (2002) Deciphering protein complexes and protein interaction networks by tandem affinity purification and mass spectrometry: analytical perspective. *Mol Cell Proteomics*, **1**, 204-212.

Staudenmann, W. and James, P. (2001) Interpreting peptide tandem mass-spectrometry fragmentation spectra. In James, P. (ed.), *Proteome Research: Mass Spectrometry*. NY: Springer.

Synder, P.A. (2000) *Interpreting Protein Mass Spectra: A Comprehensive Resource*. NY: Oxford University Press.

Tonella, L., Walsh, B.J., Sanchez, J.C., Ou, K., Wilkins, M.R., Tyler, M., Frutiger, S., Gooley, A.A., Pescaru, I., Appel, R.D., Yan, J.X., Bairoch, A., Hoogland, C., Morch, F.S., Hughes, G.J., Williams, K.L. and Hochstrasser, D.F. (1998) '98 *Escherichia coli* SWISS-2DPAGE database update. *Electrophoresis*, **19**, 1960-1971.

Tonge, R., Shaw, J., Middleton, B., Rowlinson, R., Rayner, S., Young, J., Pognan, F., Hawkins, E., Currie, I. and Davison, M. (2001) Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics*, **1**, 377-396.

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. and Rothberg, J.M. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623-627.

Uljon, S.N., Mazzarelli, L. and Chait, B.T. (2000) Analysis of proteins and peptides directly from biological fluids by immunoprecipitation/mass spectrometry. In Chapman, J.R. (ed.), *Mass Spectrometry of Proteins and Peptides*. NJ: Humana Press.

Wagner, K., Miliotis, T., Marko-Varga, G., Bischoff, R. and Unger, K.K. (2002) An automated on-line multidimensional HPLC system for protein and peptide mapping with integrated sample preparation. *Anal Chem*, **74**, 809-820.

Washburn, M.P., Wolters, D. and Yates, J.R., III. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, **19**, 242-247.

Wilkins, M.R., Gasteiger, E., Bairoch, A., Sanchez, J.C., Williams, K.L., Appel, R.D. and Hochstrasser, D.F. (1999) Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol*, **112**, 531-552.

Wilkins, M.R., Gasteiger, E., Wheeler, C.H., Lindskog, I., Sanchez, J.C., Bairoch, A., Appel, R.D., Dunn, M.J. and Hochstrasser, D.F. (1998) Multiple parameter

cross-species protein identification using MultiIdent-a world-wide web accessible tool. *Electrophoresis*, **19**, 3199-3206.

Wu, C.C. and MacCoss, M.J. (2002) Shotgun proteomics: tools for the analysis of complex biological systems. *Curr Opin Mol Ther*, **4**, 242-250.

Yamagata, A., Kristensen, D.B., Takeda, Y., Miyamoto, Y., Okada, K., Inamatsu, M. and Yoshizato, K. (2002) Mapping of phosphorylated proteins on two-dimensional polyacrylamide gels using protein phosphatase. *Proteomics*, **2**, 1267-1276.

Yan, J.X., Devenish, A.T., Wait, R., Stone, T., Lewis, S. and Fowler, S. (2002) Fluorescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of *Escherichia coli*. *Proteomics*, **2**, 1682-1698.

Yarmush, M.L. and Jayaraman, A. (2002) Advances in proteomic technologies. *Annu Rev Biomed Eng*, **4**, 349-373.

Yates, J.R., III, Carmack, E., Hays, L., Link, A.J. and Eng, J.K. (1999) Automated protein identification using microcolumn liquid chromatography-tandem mass spectrometry. *Methods Mol Biol*, **112**, 553-569.

Yates, J.R., III, Eng, J.K., McCormack, A.L. and Schieltz, D. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem*, **67**, 1426-1436.

Yates, J.R., III, Link, A.J. and Schieltz, D. (2000) Direct analysis of proteins in mixtures: application to protein complexes. In Chapman, J.R. (ed.), *Mass Spectrometry of Proteins and Peptides*. NJ: Humana Press.

Yates, J.R., III, Speicher, S., Griffin, P.R. and Hunkapiller, T. (1993) Peptide mass maps - a highly informative approach to protein identification. *Anal. Biochem*, **214**, 397-408.

Zhang, W. and Chait, B.T. (2000) ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem*, **72**, 2482-2489.

Chapter 3 Strategy for the confidence analysis of identified proteins

Part of the contents in this chapter was presented at the 3rd International Proteomics Conference (IPC'03) at Taipei (Taiwan), May 14-17, 2004. In proteomic analyses, many protein identification tools have been developed to identify proteins from mass spectra data. However, the important and most difficult task in protein identification is to achieve a high level of confidence for the searched results, even though many statistical methodologies have been employed in the protein identification tools. In our laboratory, we have developed and implemented a two-step strategy to analyze the confidence of identified proteins and/or locate results with relatively high confidence. Firstly, we used protein sequence information from the interested species and the characteristic of unique peptides to group the identified proteins into two different levels. Secondly, we cross-compared the proteins identified by different protein identification tools based on the same mass spectral data to further determine the confidence of identified proteins, e.g., common proteins have relatively high confidence. To demonstrate the strategy, two widely used protein identification packages (SEQUEST and Mascot) were used to identify proteins from the publicly accessible mass spectral data, and then the identified proteins from these tools were analyzed using the two-step strategy. The chapter was prepared in manuscript format for *Proteomics*.

3.1 Introduction

Mass spectrometry-based protein identification experiments have been the major means for large-scale proteomic studies of a cell or an organism (Link *et al*., 1999; Mawuenyega *et al*., 2003; McCormack *et al*., 1997; Peng *et al*., 2003; Pflieger *et al*., 2002). Currently there are several protein identification packages available such as MS-Tag, Mascot, SEQUEST, etc. The general procedures to interpret MS/MS spectra can be subdivided into three steps: (1) searching for the peptide sequences based on measured m/z values of precursor ions; (2) locating the most probable peptide(s) from the candidate peptides; and (3) identifying protein(s) by correlating those most probable peptide sequences with the protein sequence database. As mentioned in Section 2.4, multiple candidates are generally obtained after Step 1. Therefore, the crucial steps in MS-based protein identification include 1) how to evaluate the searched peptide results, and 2) how to correlate the searched peptide(s) to proteins. The solution methods involved in the various search tools are different.

3.1.1 Protein identification using MS-Tag

MS-Tag is a publicly accessible protein identification tool for MS and/or MS/MS spectral data. The tool was developed by Clauser *et al*. (1999) and can be accessed at http://prospector.ucsf.edu/. By providing the mass of precursor ions, the masses of the precursor ion's corresponding product ions, and pre-set search parameters such as the protein database, the tolerance of precursor ion and product ions, the charge state of searched ions, and so on, a result summary file can be generated in HTML format. The file contains not only the input data, but also the searched peptide sequences. Typically a

series of MS and MS/MS data from the same sample (e.g., from the same peptide fraction) is inputted into MS-Tag simultaneously as a batch search. Correspondingly, a complete summary report showing all identified peptides was generated and differentiated by a subtitle ‘Data Set ## Results’. For example, ‘Data Set 80 Results’ in Figure 3.1 indicates that the report part following this line is the summary related to the 80^{th} precursor ion (m/z = 591.2982 Da).

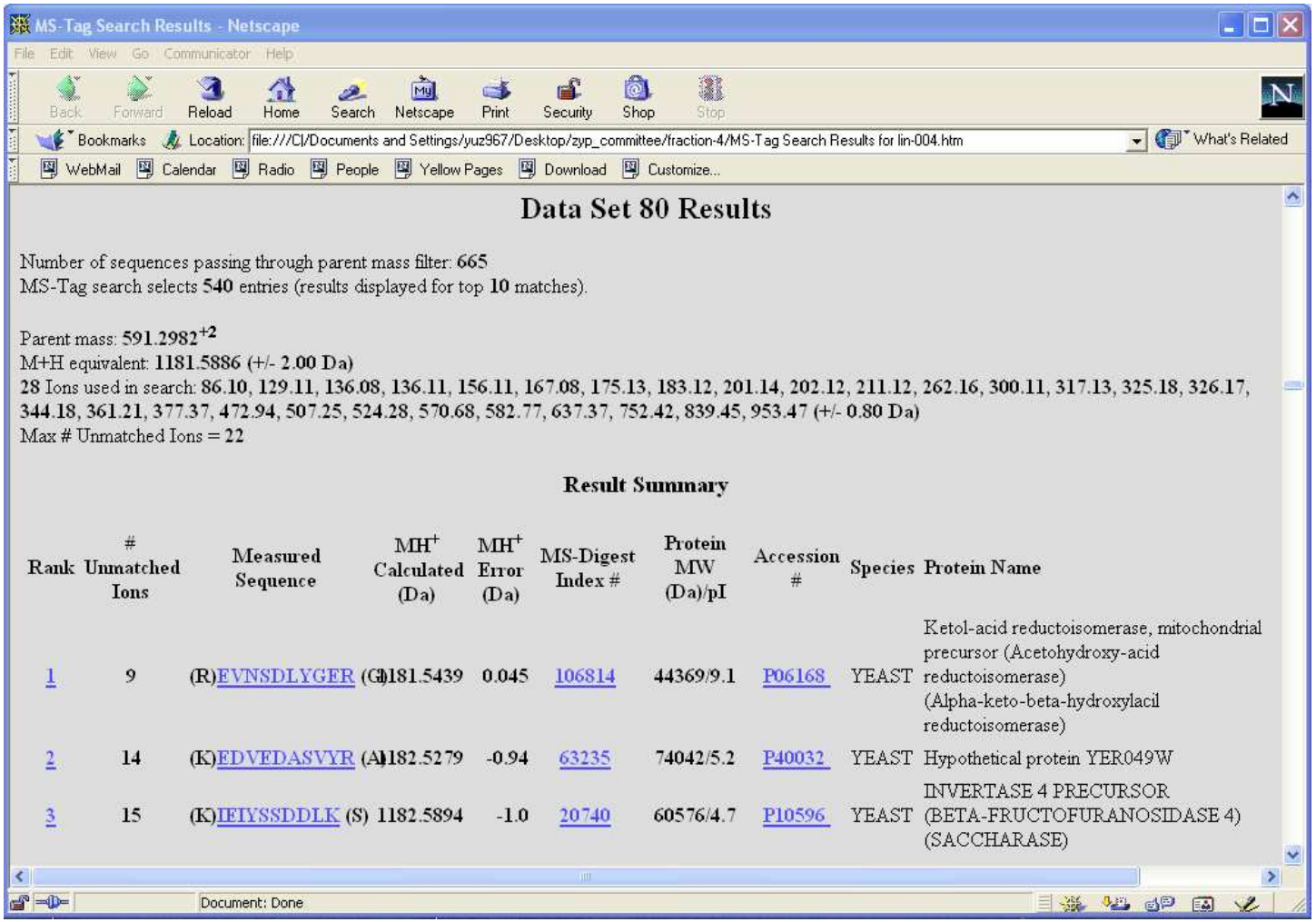


Figure 3.1 Partial listing of an MS-Tag report

It is a common phenomenon in protein identification that a set of MS data can be used to predict several peptides. For example, Figure 3.1 shows that several peptides are predicted for the same precursor ion of 591.2982 Da, so one must determine which one is real. MS-Tag does not give an answer for how to examine the search results; instead they leave the cumbersome curating work to biological researchers or MS scientists. Based on the common sense used in many software tools, the peptide ranked first is always considered the 'real' peptide, indicating that most of the experimental MS/MS peaks were matched to this peptide. For example, the sequence 'EVNSDLYGER', corresponded to a precursor ion of 591.2982 Da (the only peptide with the rank of 1) and it was regarded as the determined peptide. As a result, Protein P06168 was identified by MS-Tag.

3.1.2 Protein identification using Mascot

Like MS-Tag, a summary report was generated and presented in HTML format after a Mascot search (Perkins *et al*., 1999). The HTML file is composed of three parts: 1) summary of identified proteins (Figure 3.2); 2) summary of identified peptides (Figure 3.3); and 3) summary of un-assigned peptides (Figure 3.4).

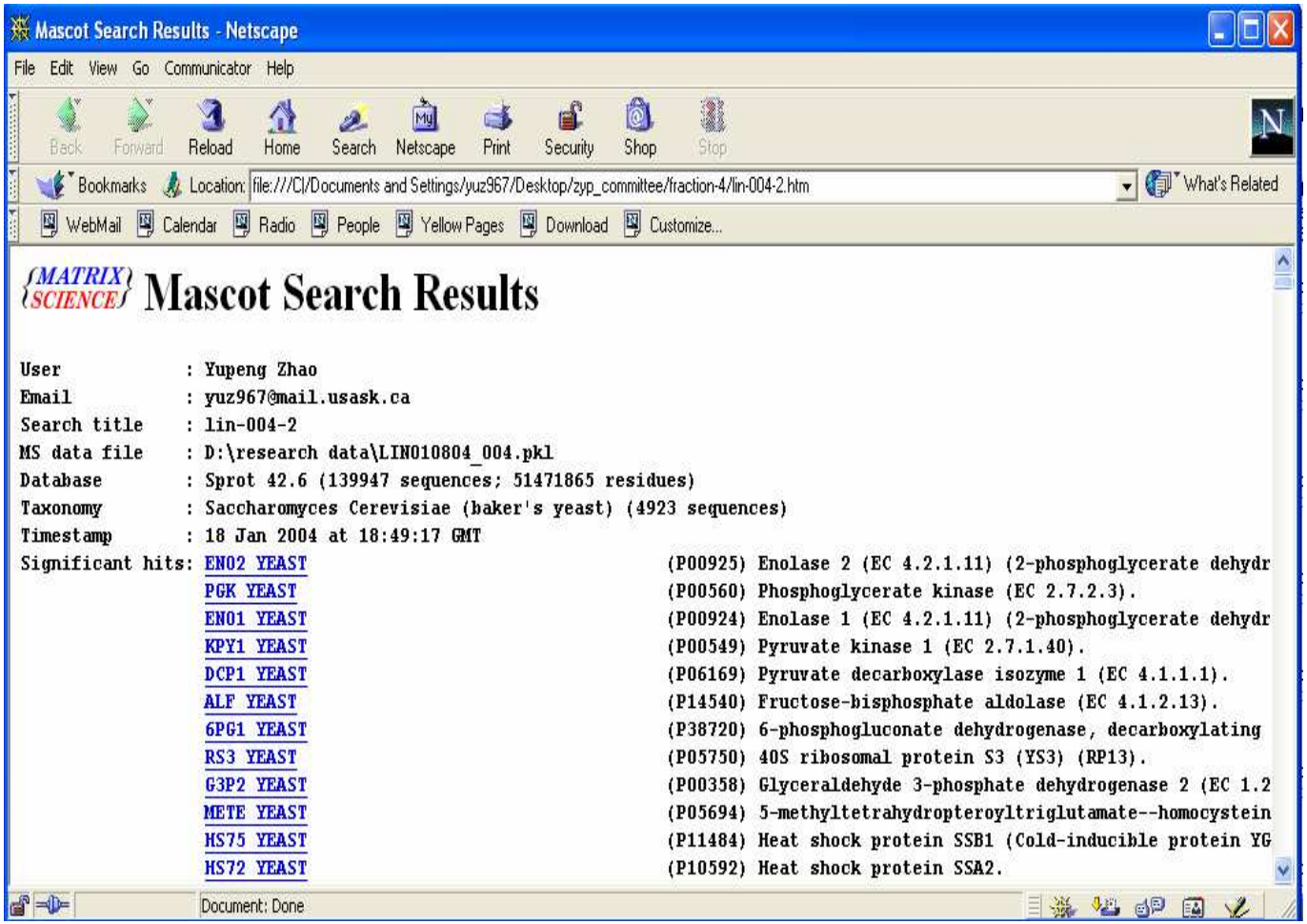


Figure 3.2 Partial listing of a Mascot protein summary report

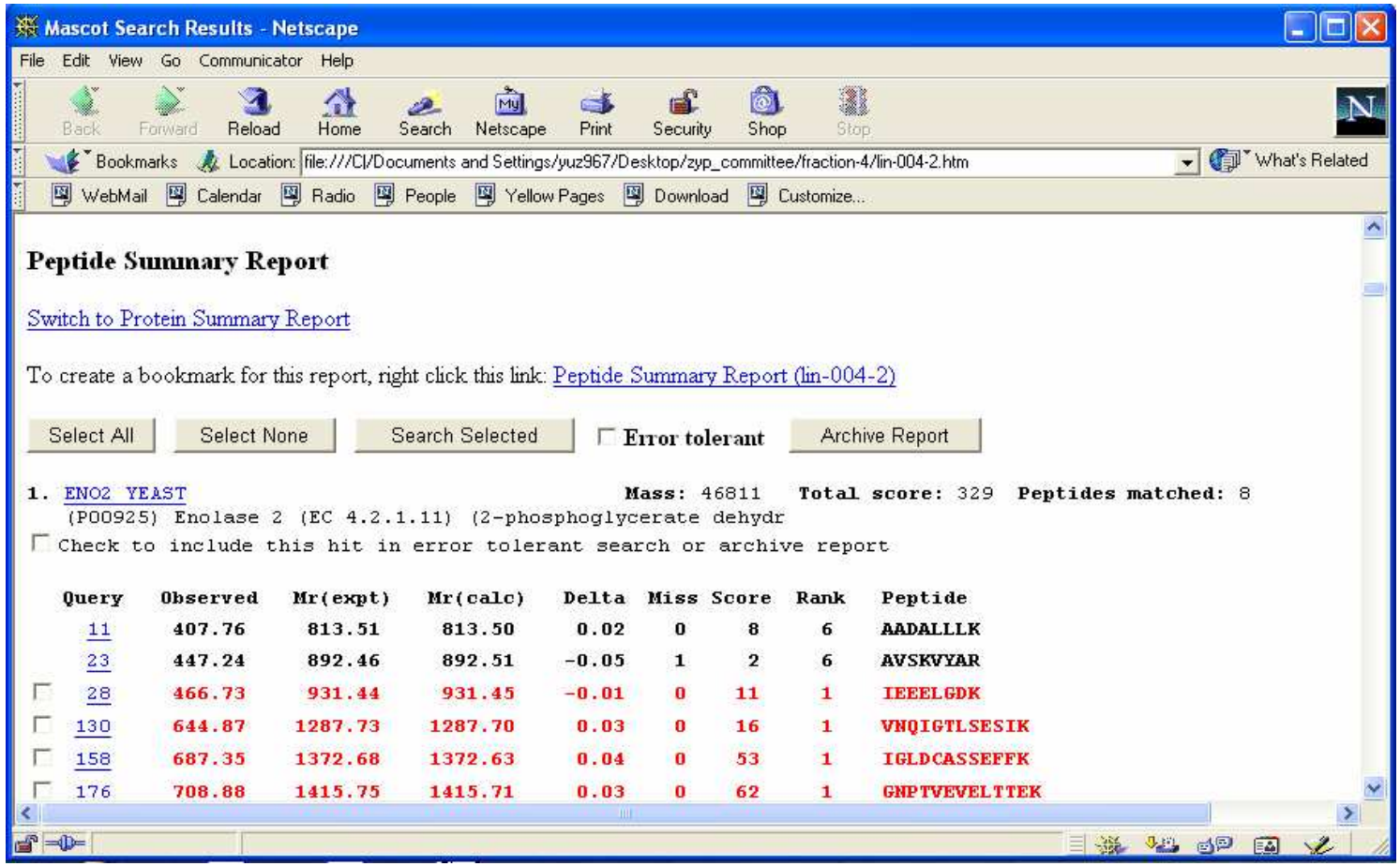
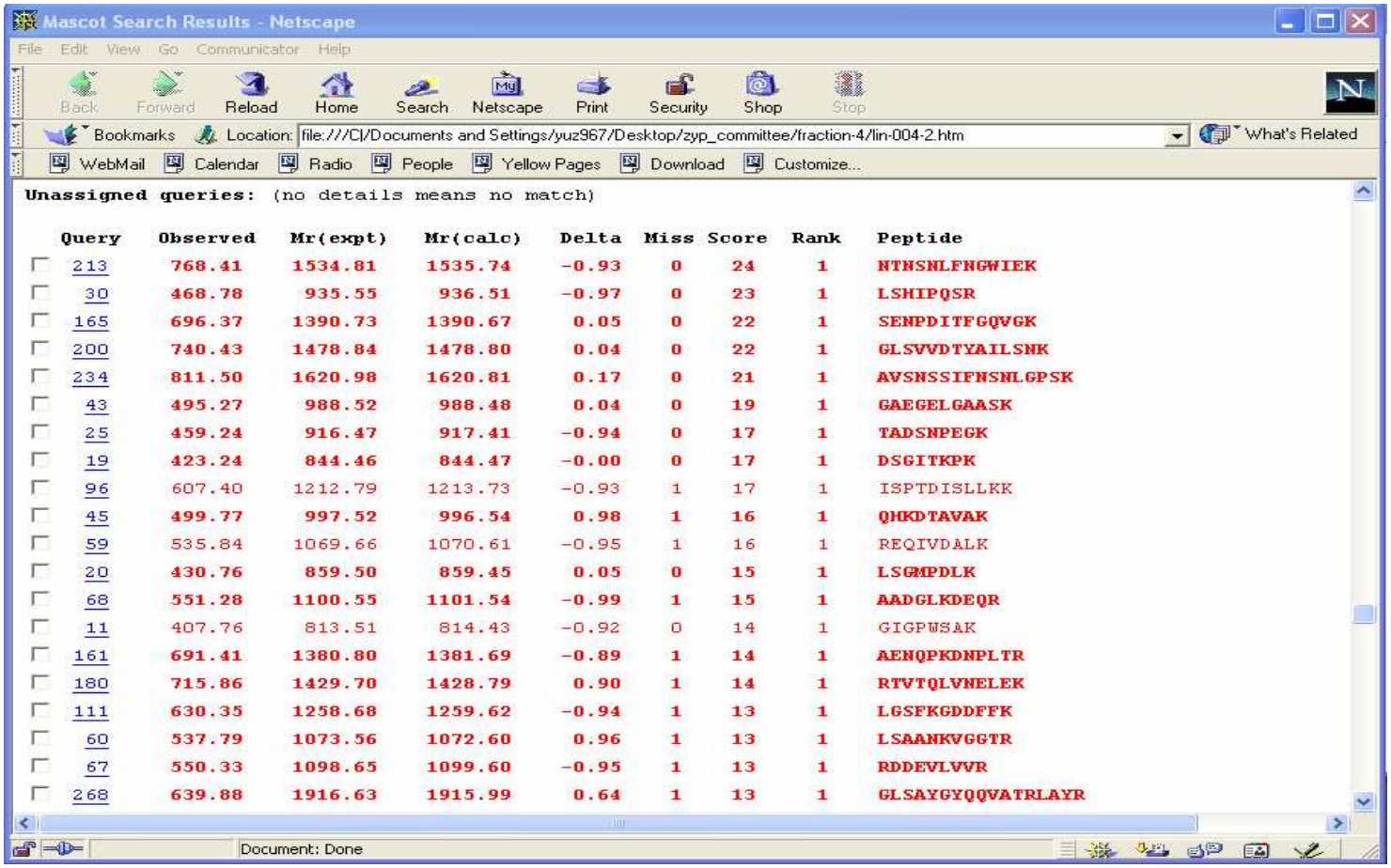


Figure 3.3 Partial listing of a Mascot peptide summary report



Unassigned queries: (no details means no match)

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Peptide
213	768.41	1534.81	1535.74	-0.93	0	24	1	NTNSNLFNGWIEK
30	468.78	935.55	936.51	-0.97	0	23	1	LSHIPQSR
165	696.37	1390.73	1390.67	0.05	0	22	1	SENPDITFGQVGK
200	740.43	1478.84	1478.80	0.04	0	22	1	GLSVVDTYAILSNK
234	811.50	1620.98	1620.81	0.17	0	21	1	AVSNSSIFNSNLGPSK
43	495.27	988.52	988.48	0.04	0	19	1	GAEGELGAASK
25	459.24	916.47	917.41	-0.94	0	17	1	TADSNPEGK
19	423.24	844.46	844.47	-0.00	0	17	1	DSGITKPK
96	607.40	1212.79	1213.73	-0.93	1	17	1	ISPTDISLLKK
45	499.77	997.52	996.54	0.98	1	16	1	QHKDTAVAK
59	535.84	1069.66	1070.61	-0.95	1	16	1	REQIVDALK
20	430.76	859.50	859.45	0.05	0	15	1	LSGMPDLK
68	551.28	1100.55	1101.54	-0.99	1	15	1	AADGLKDEQR
11	407.76	813.51	814.43	-0.92	0	14	1	GIGPWSAK
161	691.41	1380.80	1381.69	-0.89	1	14	1	AENQPKDNPLTR
180	715.86	1429.70	1428.79	0.90	1	14	1	RTVTQLVNELEK
111	630.35	1258.68	1259.62	-0.94	1	13	1	LGSFKGDDFFK
60	537.79	1073.56	1072.60	0.96	1	13	1	LSAANKVGGTR
67	550.33	1098.65	1099.60	-0.95	1	13	1	RDDEVLVVR
268	639.88	1916.63	1915.99	0.64	1	13	1	GLSAYGYQQVATRLAYR

Figure 3.4 Partial listing of a Mascot un-assigned peptide summary report

Mascot ranks the peptide candidates in a decreasing list based on the calculated probability (Mowse score) for each peptide. The Mowse score indicates the confidence of identification of an identified peptide. A threshold value can be used as a judgement point, and a peptide with a Mowse score greater than the threshold value is considered as a confident identification.

When correlating the identified peptides to proteins, Mowse uses a method like ‘sequence coverage’ (see Section 2.5.3 for definition). That is, all Mowse scores of peptides belonging to one protein are added, and this total score is used as a final score for the protein. If the final score is greater than the threshold value, this protein is considered as identified. Figure 3.5, for example, demonstrates the ‘sequence coverage’ application in Mascot. For Protein P14540 (gene name: ALF_YEAST), two peptides were identified with a total score of 160, which is greater than 25 (25 is a threshold value determined by Mascot basing on given searching criteria), such that this protein is reported as identified; the same conclusion is drawn for protein P38720. According to this method, it is also possible that a protein can be interpreted using just one identified peptide under the condition that the Mowse score of this identified peptide greater than the threshold Mowse value (Figure 3.6).

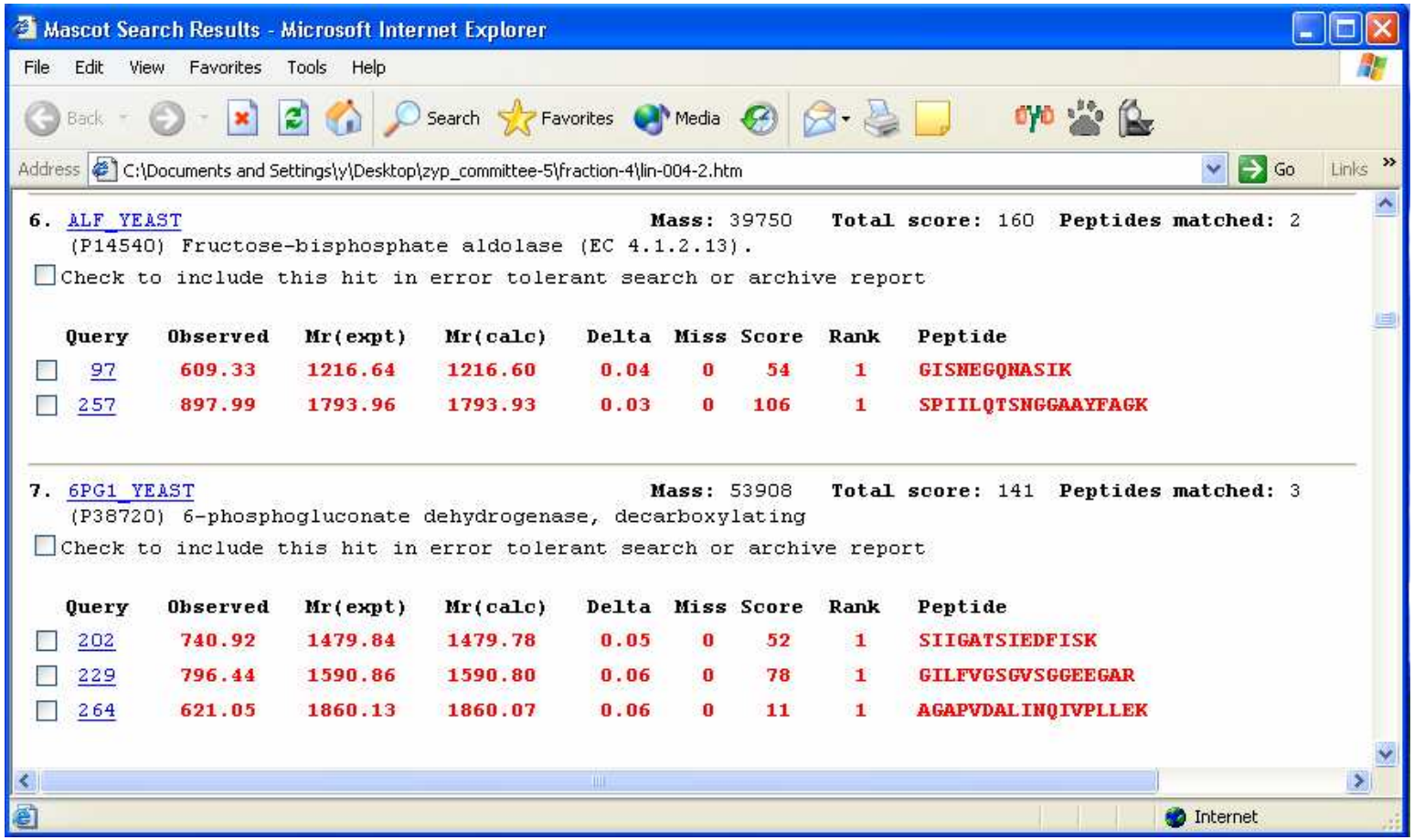


Figure 3.5 Sequence coverage for protein identification

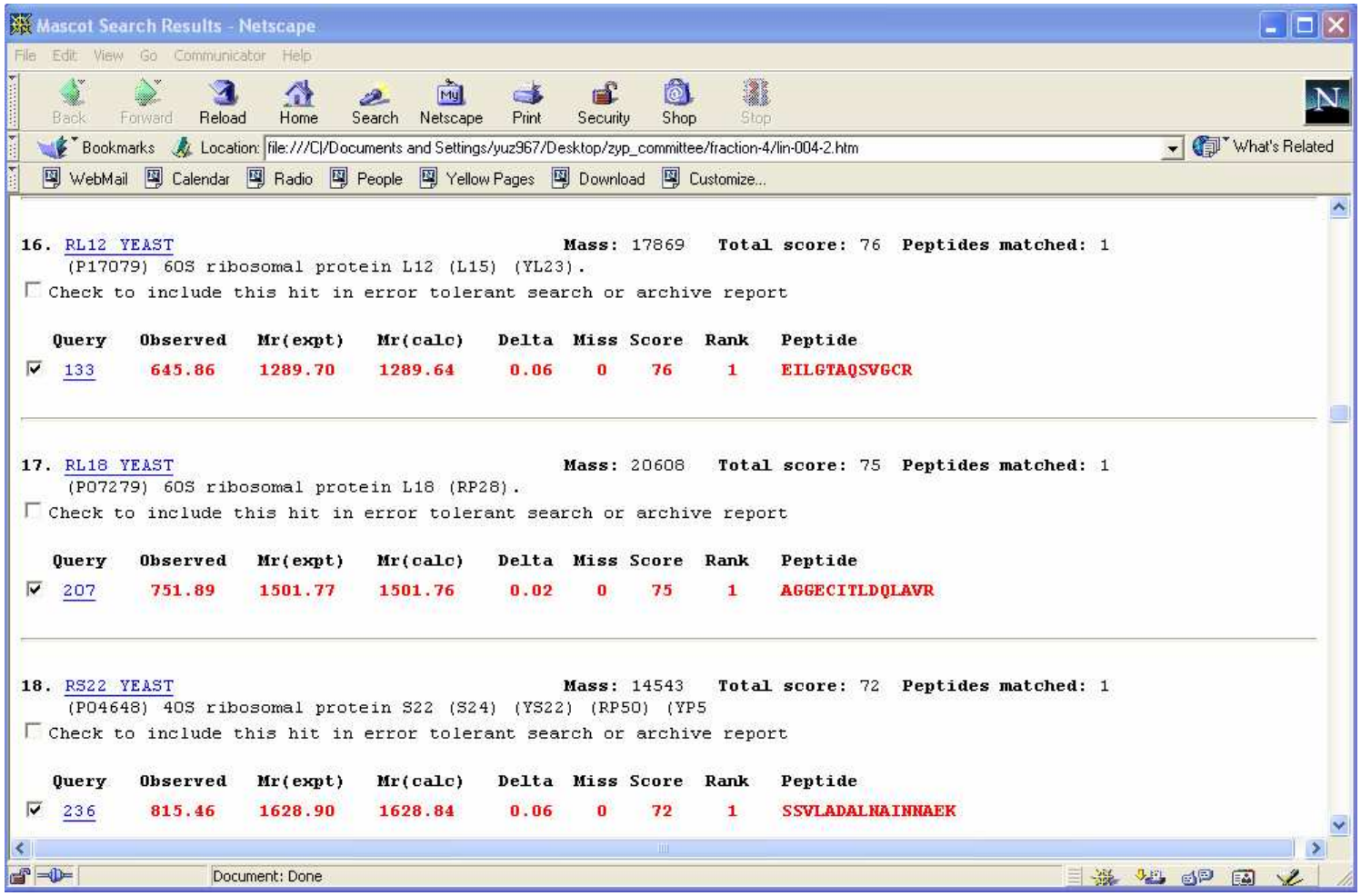


Figure 3.6 Proteins predicted from just one peptide

3.1.3 Protein identification using SEQUEST

In a different fashion from MS-Tag and Mascot, SEQUEST (Eng *et al*., 1994; Yates *et al*., 1995) searches against a protein database using one set of MS and MS/MS data in each run, and then the searched peptides are ranked in a decreasing list based on the calculated X_{corr} value through cross-correlation analysis (Figure 3.7). Therefore, many output files are generated for a sample fraction. To analyze the searched results, SEQUEST uses pre-set threshold values to determine the 'real' peptide. The peptide that ranked first and considered as a 'real' peptide should have an X_{corr} value greater than the threshold value (e.g., 2.0 or 2.5).

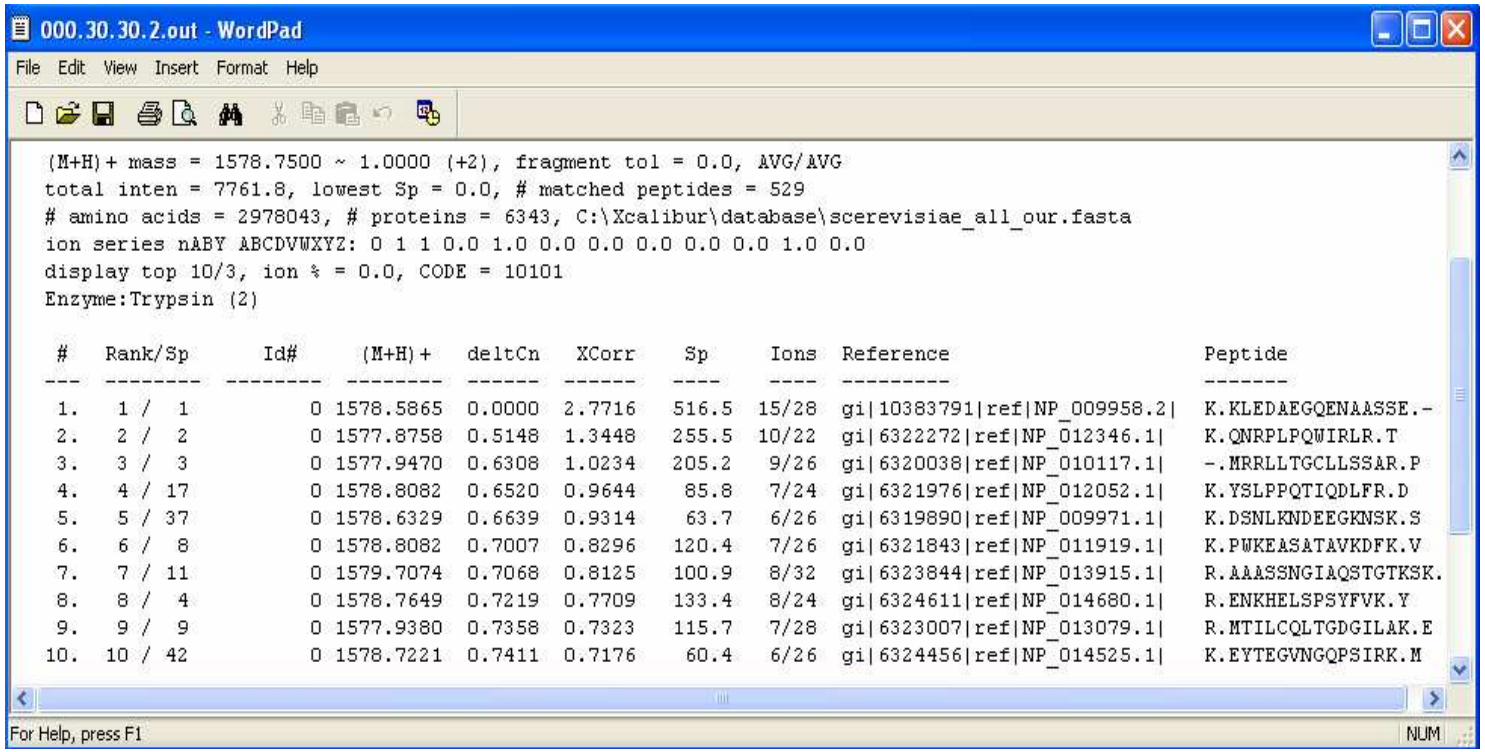


Figure 3.7 Partial listing of a SEQUEST report

3.1.4 Commentary for protein identification tools

It is not easy to differentiate which protein identification strategy is better. MS-Tag is simple and works well for high-quality experimental data. Generally, the matched peptide sequence(s) with a rank of 1 is regarded as the identified peptide, indicating that most of the experimental MS/MS peaks were matched to this peptide(s). For example, the sequence ‘EVNSDLYGER’ corresponding to the precursor ion of 591.2982 Da and the sequence ‘VVDLIEYVAKA’ interpreted from the precursor ion of 610.3668 Da (Figure 3.8) were regarded as identified, correspondingly, the proteins P06168 and P00360 were identified.

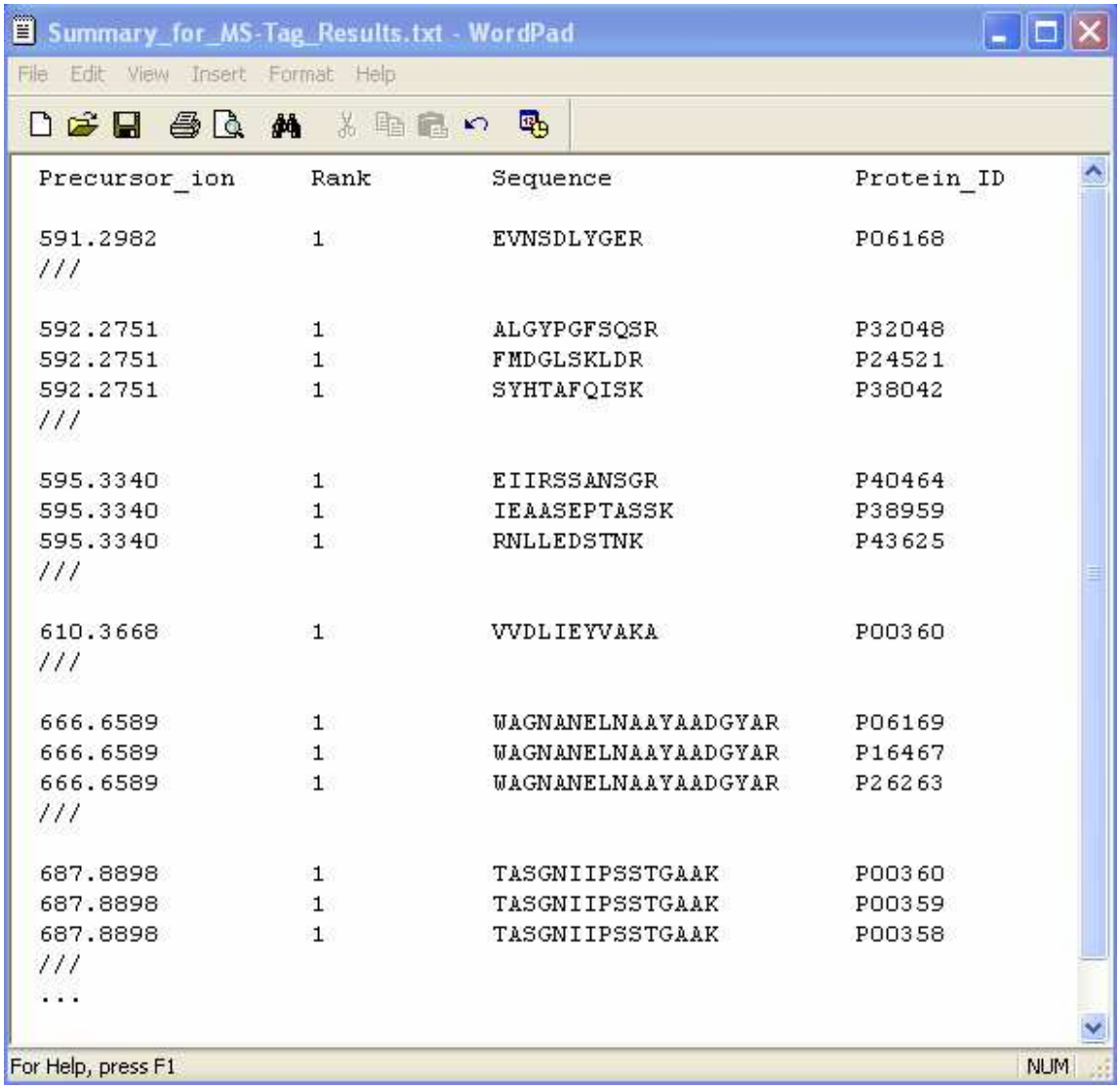


Figure 3.8 Partial listing of an MS-Tag summary report. There are 3 situations shown in this report. Firstly, for a precursor ion of 591.2982 Da, only one peptide is identified and ranked first, such that its corresponding protein (P06168) is deemed identified. This reasoning is also applicable to the precursor ion of 610.3668 Da. Secondly, for a precursor ion of 666.6589 Da, three identical peptide sequences result in the identification of 3 different proteins (P06169, P16467, and P26263), leading to a questionable conclusion. A similar interpretation can also be applied to the precursor ion of 687.8898 Da. Finally, for the precursor ion of 592.2751 Da, three different proteins (P32048, P24521, and P38042) result from 3 distinct peptide sequences; hence no deterministic deduction can be drawn. The above logic can also be applied to the precursor ion of 595.3340 Da.

However, there are two problems associated with MS-Tag. The first one is that two or more peptides having the same sequence and the same rank (e.g., rank = 1) can be identified. For instance, three of the same sequences 'WAGNANELNAAYAADGYAR' for the precursor ion of 666.6589 Da and three same sequences 'TASGNIIPSSTGAAK' for the precursor ion of 687.8898 Da were identified (Figure 3.8). The reason for the same sequence output by MS-Tag is that these three sequences were from different proteins in the target protein database, and MS-Tag searches proteins one by one in the database. Under this condition, all corresponding proteins are considered identified, i.e., proteins P06169, P16467, and P26263 for the precursor ion of 666.6589 Da and proteins P00360, P00359, and P00358 for the precursor ion of 687.8898 Da. Obviously, there are questions about such proteins. These proteins are not confidently identified. The second problem of using MS-Tag is when two or more different peptide sequences with the rank of 1 were located for the same precursor ion, i.e., the sequences for precursor ion of 592.2751 Da and 595.3340 Da as shown in Figure 3.8. Generally these peptides are considered as un-identified in this case, because it is almost impossible to decide which one was the 'real' one.

The obvious disadvantage in protein identification using SEQUEST is how to determine the pre-set threshold value (X_{corr}). Under different X_{corr} settings, the searched results based on the same MS/MS data may vary greatly, bringing confusion to biological researchers. For example, from the MS/MS data of *S. cerevisiae* (Prince *et al*., 2004), 1227 proteins was identified when X_{corr} was set to 2.0 or greater, whereas only 347 proteins were identified when X_{corr} was set to greater than or equal to 2.5. These two sets of "identified" proteins were derived from the same MS/MS spectral data set using the

same software tool; however, a large difference between groups of identified results was obtained. Thus these protein results will bring confusion for biological researchers when applying them to interpret phenotypic observations. Conflicting conclusions might even be drawn.

As for Mascot, a wrong protein prediction is possible under the following two conditions. The first one is shown in Figure 3.9. Even though three peptides were identified using Mascot and the total Mowse score was 27 (greater than the threshold Mowse score, 25), Protein Q07807 was not considered as a confident identification because each peptide identified here had a low Mowse score.

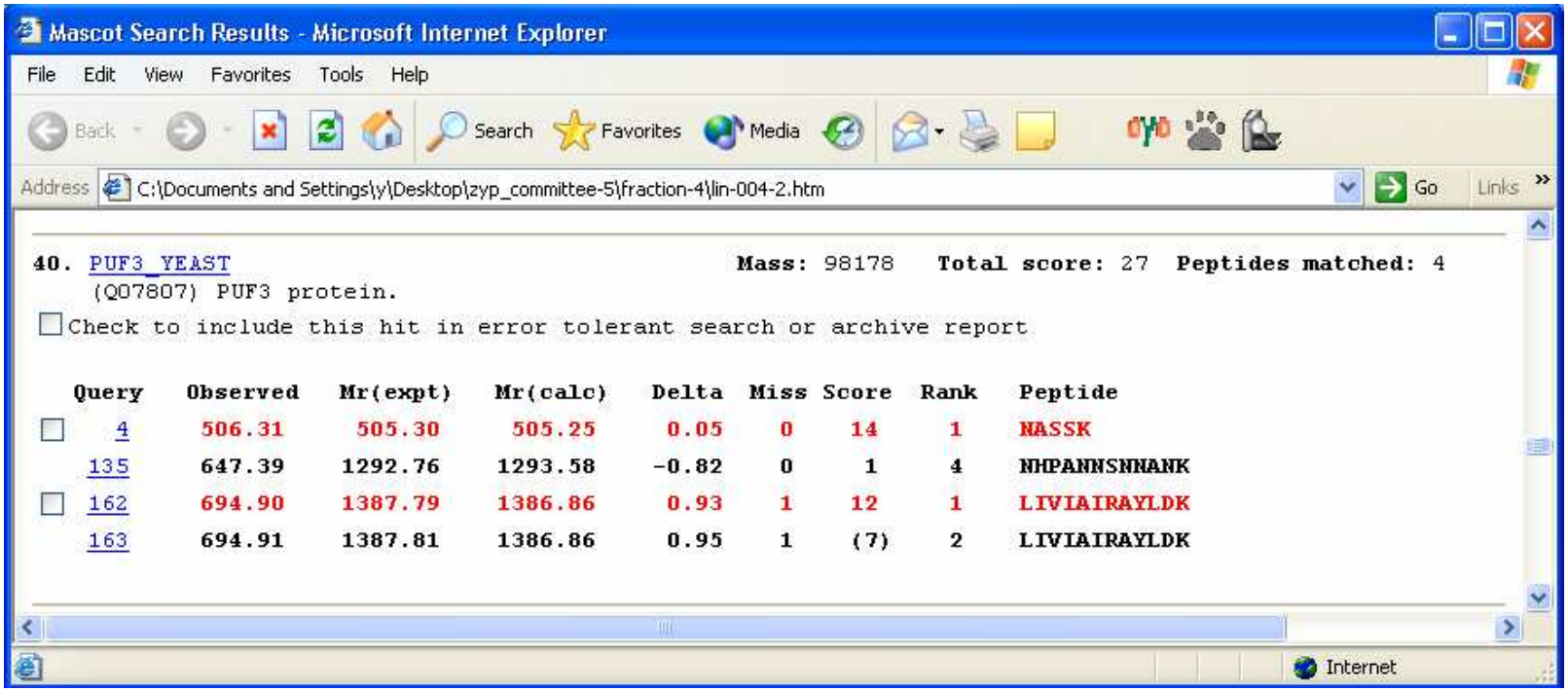


Figure 3.9 Protein predicted from peptides with low probability

The second condition is shown in Figure 3.10. The identified peptide(s) with high Mowse score may be applied to predict several proteins because the identified peptide sequence(s) exist in several proteins. For example, proteins P40439, P53341, P38158, and P07265 were derived from the same peptide sequence. Therefore, it is difficult to deduce which protein really exists. Some researchers (e.g., Mawuenyega, *et al*., 2002) have assumed that all of these proteins were identified under this condition; however, this may lead to wrong conclusion(s) for subsequent research activities.

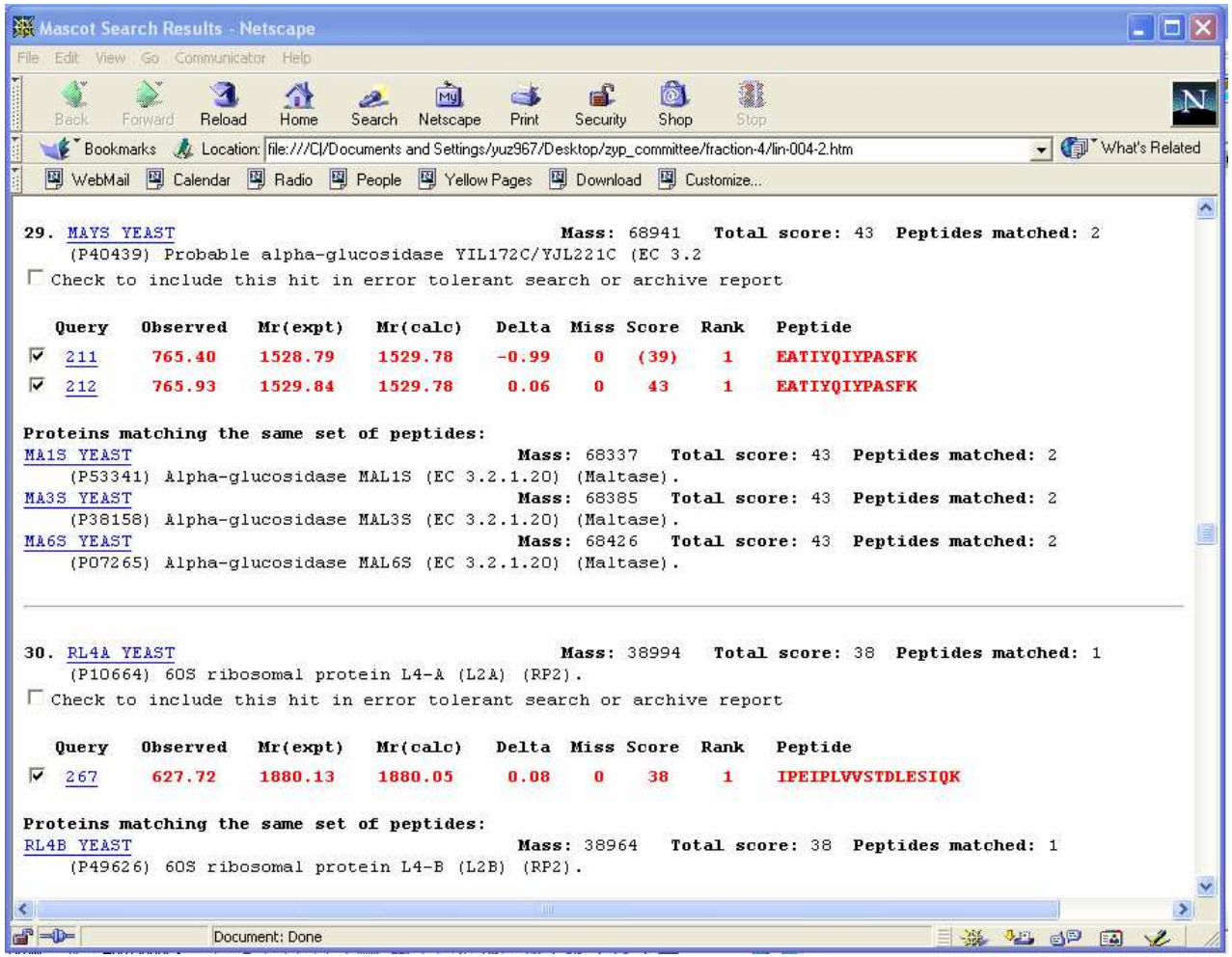


Figure 3.10 Several proteins predicted from the same identified peptide(s)

3.2 Strategy for the confidence analysis of identified proteins

Confidence, describing the probable level of identification of the identified proteins, is an important issue for proteomic research. A set of highly confident results can lead to accurate conclusions. The disadvantages involved in different tools lead us to question the protein results identified if only a single identification tool is used. Chamrad *et al*. (2004) applied different protein identification tools to the same set of MS and MS/MS spectral data and observed that only 30-50% of results were consistent. This implies that the searched proteins from each protein identification tool have different confidences, and only those proteins with high confidences can be identified by different tools. Thus a strategy to analyze the confidences of searched proteins is really in need.

In our laboratory, a two-step approach to analyze the confidence of identified proteins was developed based on the unique peptide concept and cross comparison (Figure 3.11). The unique peptide concept can analyze the confidences of searched proteins by a single identification tool, while the cross comparison can help locate the high confidence proteins by looking for the common proteins identified by the different identification tools. Two steps involved in this approach can be applied separately or in a combined mode, depending on the availability of extra protein identification tools. The detailed information of this two-step approach is illustrated in the following sections.

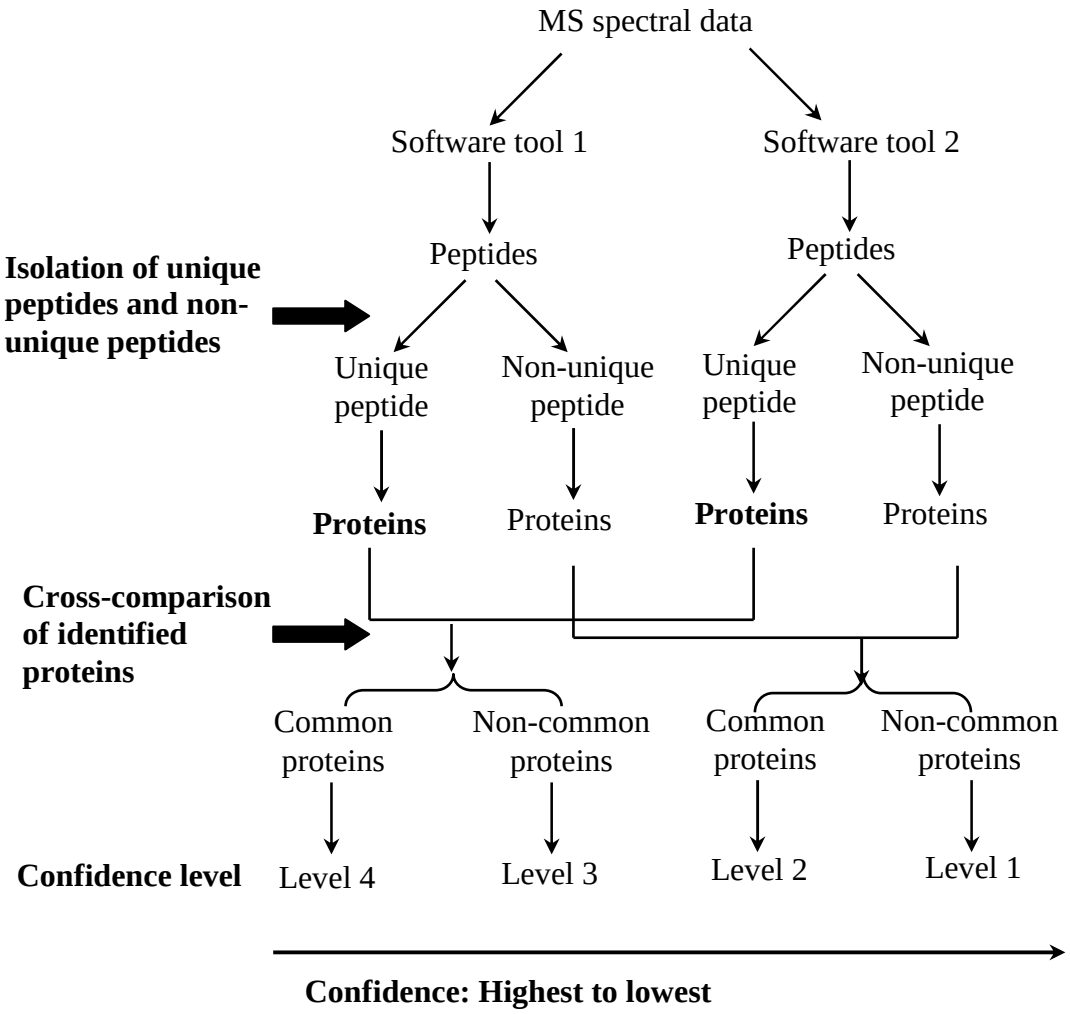


Figure 3.11 Schematic of strategy to analyze protein identification confidence

3.2.1 Unique peptide

A unique peptide is defined as a peptide that exists in only one protein in the protein pools of interest, although this peptide may appear more than once in the same protein. For example, assuming that Proteins 1 and 2 are digested by trypsin, and the generated peptides are listed in Figure 3.12, respectively.

ANDR NQEGHK **MFPSTK** WYVTR NQEGHK (Protein 1)

CEGIK MFPSR WYVTR **MFPSTK** CEGIK (Protein 2)

Figure 3.12 Amino acid sequences of pseudo proteins 1 and 2
Peptides generated from Protein 1:
ANDR, NQEGHK, MFPSTK, WYVTR, NQEGHK
Peptides generated from Protein 2:
CEGIK, MFPSR, WYVTR, MFPSTK, CEGIK

By our definition, the peptide ANDR shown in Figure 3.12 is regarded as a unique peptide because it appears once in Protein 1 but not at all in Protein 2. The peptide NQEGHK is also considered unique because it is not observed in Protein 2 although it appears twice in Protein 1. Neither MFPSTK nor WYVTR are unique peptides as they appear in both proteins 1 and 2. Other unique peptides are MFPSR and CEGIK, only found in Protein 2.

The definition of ‘unique peptide’ is essential in protein identification. For example, it is straightforward to identify Protein 1 if either ANDR or NQEGHK, or both are identified, whereas it is difficult to conclude whether Protein 1 or 2 exist if only MFPSTK is identified from the MS/MS data. Therefore, a unique peptide can act as a ‘protein-tag’ in protein identification.

In our proposed two-step approach, the first step is to analyze the confidence of identified proteins by a single identification tool. During the protein identification, a protein identification tool firstly identifies the ‘real’ peptides from the experimental mass spectral data and then predicts the proteins from these ‘real’ peptides. In our approach, the ‘real’ peptides firstly are divided into unique peptides and non unique peptides; then the proteins identified from one or more unique peptides are grouped into highly confident proteins and the proteins identified from non unique peptides are considered as low confident ones.

The proposed unique peptide concept is applicable since unique peptides are distributed largely in the trypsin-treated peptide pool (Figure 3.13). There are total 364,864 unique peptides in our *S. cerevisiae* database, representing 84.6% of the total peptides (431,041). In the mass range of 400-4800 Da, which is a typical MS scan range for precursor ions in proteomic studies, the unique peptides can be applied to identify 4,896 proteins, representing 99.4% of the total proteins of *S. cerevisiae* (4,923). This makes it possible to apply unique peptide definition in the protein identification.

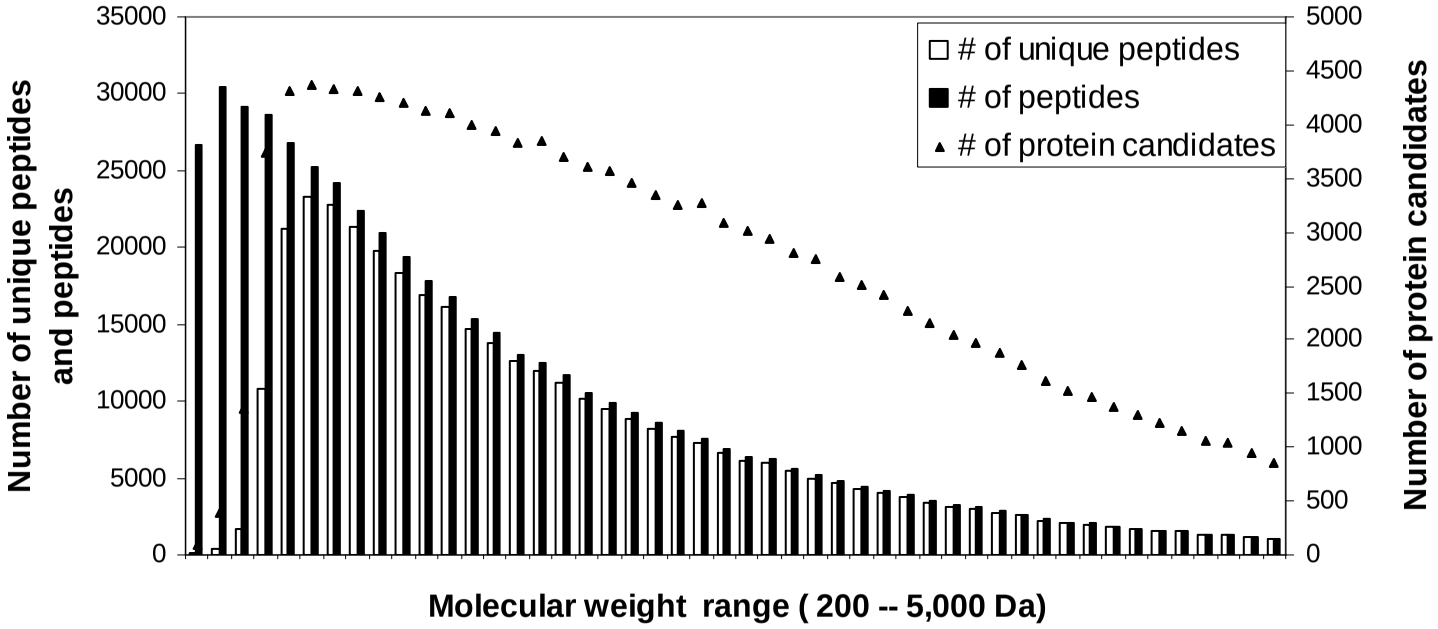


Figure 3.13 Distributions of unique peptides, peptides, and protein candidates

3.2.2 Cross comparison

It is common that different peptides are identified when the same MS and MS/MS spectral data are interpreted by different software tools (Chamrad *et al*., 2004). Therefore, different proteins might be reported by various identification tools, even though the unique peptide concept is applied in the tools to help analyze the confidence of identified proteins. Cross comparison can improve the confidence by finding the common proteins in different protein identification tools.

The second step in our developed strategy is, therefore, cross comparison to further analyze the confidence of searched proteins among different protein identification tools from the same mass spectral data. The common proteins from all tools are considered as highly confident results, while the others are low confident results. This process is refereed to as cross comparison. No proteomic project has been reported using this approach.

After the two-step analysis, the identified proteins can be grouped into four groups with different levels of confidence. The proteins identified by the unique peptide concept and found from different protein identification tools in the cross comparison are grouped into level 4, the group with the highest confidence. The proteins identified from unique peptides that do not pass the cross comparison test are grouped into level 3. The proteins identified from non-unique peptides that do pass the cross comparison test are grouped into level 2. Finally, the proteins identified from non-unique peptides that do not pass

the cross comparison test are grouped into level 1, the group identified with the lowest confidence.

3.3 Implementation

The proposed approach was tested in connection with experimental data retrieved from http://bioinformatics.icmb.utexas.edu/OPD/. Prince *et al*. (2004) and their coworkers carried out several LC-MS/MS analyses for many different kinds of species (e.g., *E. coli*, *S. cerevisiae*, Human cell lines, and so on). They have compiled and posted mass spectral information and the searched results (using SEQUEST) on the above website for public applications, e.g., providing MS data for programmers to check their algorithms. We retrieved one sample set of MS/MS spectral data (11 fractions in total; Organism: *Saccharomyces cerevisiae*, Acc#: opd00034_YEAST, and Name: 6-04-03-YPD_test) and the corresponding protein set (seqsum.zip) identified using SEQUEST for the purpose of illustration. These MS spectral data were also fed into MS-Tag and Mascot for protein interpretation. However, only the searched results from SEQUEST and Mascot were analyzed and compiled for the demonstration purpose because 1) these two software tools are the most widely used commercial packages; and 2) two software tools are considered sufficient to demonstrate the proposed two-step strategy. Figure 3.14 shows the final results of using our proposed two-step strategy.

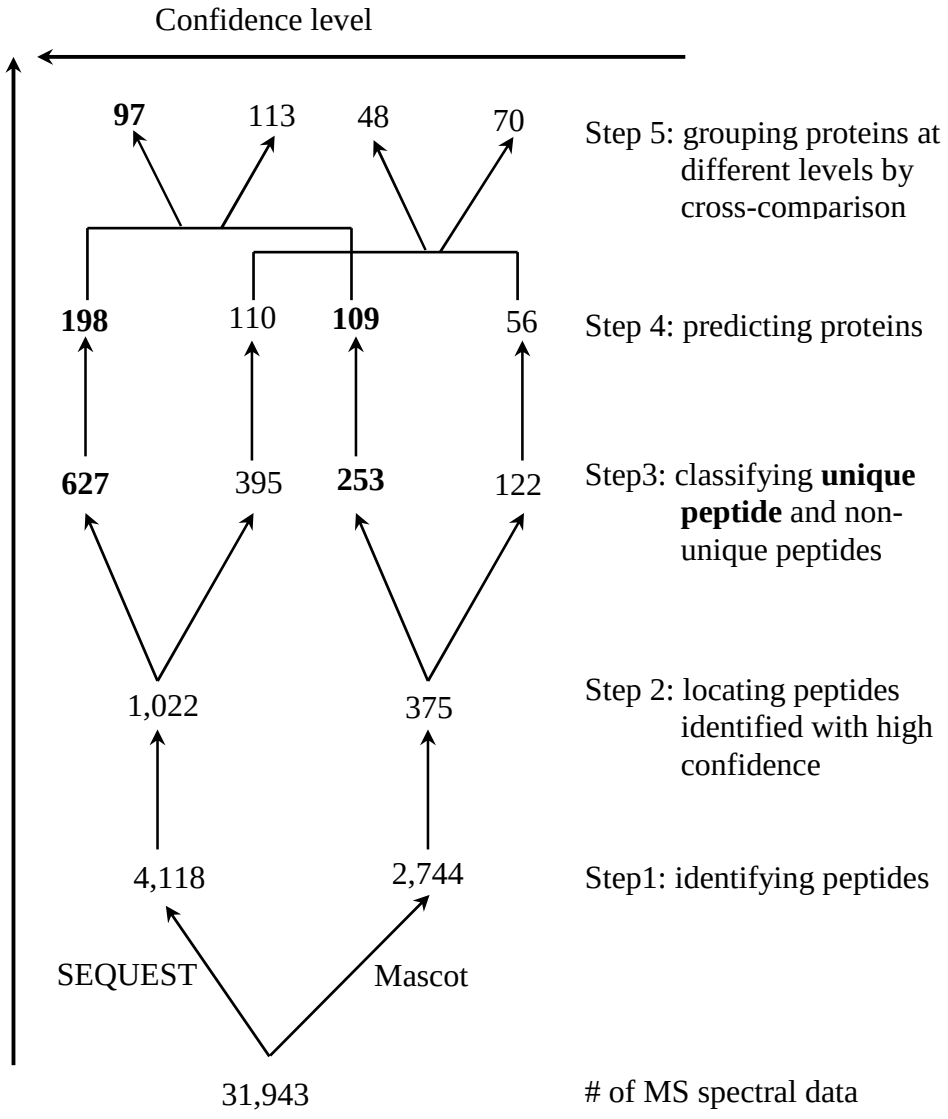


Figure 3.14 Proteins identified with various degrees of confidence using the proposed two-step strategy

A traditional protein identification tool contains three steps (Step 1, 2 and 4 as shown in Figure 3.14) to interpret mass spectral data. Firstly, peptides are identified based on the input mass spectral data. It is typical that not all experimental mass spectral data are useful during protein identification. For example, only 4,118 peptides were derived from 31,943 sets of MS and MS/MS spectral data using SEQUEST, while only 2,744 peptides were identified using Mascot for the same mass spectral data. A similar case was reported by Peng *et al*. (2003) in their yeast proteome experiment, where 162,000 MS/MS spectral data were generated using LC-MS/MS with a mass scan range of 400-1700 Da. Among the obtained spectral data, only 26,815 peptides were identified, representing only 16.5% of the original mass spectral data. Secondly, the identified peptides were then checked for 'confident' peptides using the accompanying criteria of each protein identification tool. For example, the threshold values for choosing confident peptides from all of the identified peptide pool are: in SEQUEST, the X_{corr} was set greater than 1.5, 2.0, and 3.3 for the peptide's charge state of +1, +2, and +3, respectively (Peng *et al*., 2003), while in Mascot, the Mowse value was set greater than 26. Using these identification criteria, 1,022 peptides from the 4,118 identified peptides using SEQUEST were considered as 'confident' results, while only 375 peptides among the Mascot results were 'confident'. Finally, proteins were identified based on the 'confident' peptides. As to Mascot, it takes into account the 'non-confident' peptides as well. If the total Mowse score of two or more of the 'non-confident' peptides that come from the same protein was greater than the threshold value, the corresponding protein was also considered identified.

The disadvantages of the traditional protein identification techniques were discussed in the Commentary section (Section 3.1.4). Our proposed two-step strategy involves two more steps (Step 3 and 5 in Figure 3.14) in addition to the traditional three-step approach.

In Step 3, the 1,022 ‘confident’ peptides from SEQUEST were divided into two groups, 627 unique peptides and 395 non-unique peptides. As a results, 198 proteins were identified using the unique peptide concept and were considered highly confident, while the 110 proteins identified from non-unique peptides were considered as low confidence. Similarly, 109 highly confident proteins were identified from 253 unique peptides using Mascot tool and 56 proteins were identified from 122 non-unique peptides (Figure 3.14). This great discrepancy in identified proteins further certifies the observation reported by Chamrad *et al*. (2004) that care should be taken when applying these searched results.

To further analyze the confident proteins obtained with different protein identification tools, the cross-comparison method, Step 5, was applied after the first step analysis (unique peptide analysis). Among the 198 confident proteins from SEQUEST and 109 confident proteins from Mascot, 97 proteins were found common and were considered as the highest confident results (Level 4). The other 113 proteins identified by SEQUEST and Mascot were considered as the second highest confident results (Level 3). Similarly, after comparing the 110 proteins from SEQUEST and 56 proteins from Mascot, 48 proteins were grouped into Level 2 and 70 proteins were grouped into Level 1, the lowest confident group.

3.4 Concluding remarks

The proposed two-step approach can group identified proteins into four levels of confidence. Therefore, researchers can apply the identified proteins according to the confidence. The number of confidently identified proteins certainly decreased greatly, e.g., only 97 proteins were considered as the highest confidence from the original 31,943 sets of mass spectral data after unique peptide analysis and cross comparison (Figure 3.14). The conclusions drawn from these proteins are considered highly confident. For the lower level confident proteins, it is recommended that the researchers carry out further or alternative experiments (i.e., 2D-PAGE, western blot, etc) to verify their existence before drawing conclusions.

3.5 References

Chamrad, D.C., Korting, G., Stuhler, K., Meyer, H.E., Klose, J. and Bluggel, M. (2004) Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*, **4**, 619-628.

Clauser, K.R., Baker, P. and Burlingame, A.L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem*, **71**, 2871-2882.

Eng, J.K., McCormack, A.L. and Yates, J.R., III. (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.

Fenyo, D. (2000) Identifying the proteome: software tools. *Curr Opin Biotechnol*, **11**, 391-395.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M. and Yates, J.R., III. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol*, **17**, 676-682.

Mawuenyega, K.G., Kaji, H., Yamuchi, Y., Shinkawa, T., Saito, H., Taoka, M., Takahashi, N. and Isobe, T. (2003) Large-scale identification of *Caenorhabditis elegans* proteins by multidimensional liquid chromatography-tandem mass spectrometry. *J Proteome Res*, **2**, 23-35.

McCormack, A.L., Schieltz, D.M., Goode, B., Yang, S., Barnes, G., Drubin, D. and Yates, J.R., III. (1997) Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal Chem*, **69**, 767-776.

Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. and 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.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551-3567.

Pflieger, D., Le Caer, J.P., Lemaire, C., Bernard, B.A., Dujardin, G. and Rossier, J. (2002) Systematic identification of mitochondrial proteins by LC-MS/MS. *Anal Chem*, **74**, 2400-2406.

Prince, J.T., Carlson, M.W., Wang, R., Lu, P. and Marcotte, E.M. (2004) The need for a public proteomics repository. *Nat Biotechnol*, **22**, 471-472.

Yates, J.R., III, Eng, J.K., McCormack, A.L. and Schieltz, D. (1995b) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem*, **67**, 1426-1436.

Chapter 4 A proteomic tool for protein identification from tandem mass spectral data

This chapter has been published in *Proteomics*, **5,** 853-855 (2005). Part of the contents was presented at the 1st Canadian Plant Genomics Workshop at Saskatoon (Canada), August 23-26, 2003.

4.1 Abstract

The development of an efficient algorithm to interpret MS and MS/MS data collected from tandem mass spectrometry has attracted much attention. The proposed two-pass approach searches a species-specific peptide database based on the experimentally obtained MS and MS/MS data. In the first pass of the approach, a species-specific peptide database is generated using publicly accessible genome information. The m/z of a precursor ion is searched against the peptide database to obtain a list of candidate peptides along with the corresponding proteins. In the following pass, the MS/MS data of fragment ions derived from the same precursor ion are used to identify the most probable protein.

Instead of using probability, a simple and yet effective heuristic approach was employed to treat experimentally obtained MS/MS data for protein identification. The proposed approach is based on the total number (T) of identified experimental MS/MS

data. To warrant the subsequent ranking, the total number of identified b- and y-type ions (T_{b+y}) must be greater than 50% of T. Peptides having the same T and T_{b+y} are either ranked by the contiguity of identified ions or discarded during identification. When compared to other protein identification tools, the searched results agreed.

4.2 Introduction

Proteomics is the study of all expressed proteins of a cell or organism grown under various conditions. The information so obtained is essential to interpret physiological characteristics, metabolic alterations, transcriptional and translational modifications, and even protein-protein interactions. Mass spectrometric analysis (MSA) has recently been the major instrument for protein identification (Aebersold and Mann, 2003; Gygi and Aebersold, 2000; Link *et al*., 1999; Peng *et al*., 2003; Peng and Gygi, 2001; Washburn *et al*., 2001).

The crucial and time-consuming step in proteomic analysis is the interpretation of MS and MS/MS data obtained from MSA. The basic logic of the interpretation consists of 1) search against a protein database of the species of interest for a list of candidate peptides according to the mass-to-charge ratio (m/z) of the selected precursor ion; and 2) identify the most probable peptide from the candidate list according to the experimental MS/MS data. Several MS and MS/MS interpretation tools that are currently available include Mascot (Perkins *et al*., 1999), PepFrag (Qin *et al*., 1997), MS-Tag (Clauser *et al*., 1999), PepSea (Mann and Wilm, 1994), and SEQUEST (Eng *et al*., 1994). Although similar logics have been implemented, different searching and scoring criteria were developed.

When scoring peptide sequences, most interpretation tools use the re-constructed MS/MS spectra for the basis of peptide-ranking. Instead, we propose to use the uninterpreted experimental MS/MS data as the reference. Additionally, several heuristic rules were defined, resulting in a simple and yet effective peptide and protein identification. The approach was compared and validated using other available tools.

4.3 Methods

4.3.1 Logic

The general strategy for interpreting MS and MS/MS data is depicted in Figure 4.1. Briefly, a protein database of the species of interest is retrieved from a publicly accessible genome web site, followed by computer-aided proteolysis by simulating trypsin; a theoretical peptide database is thus constructed. By providing experimentally obtained m/z and z values of precursor ions, a list of candidate peptides having the same MW as the query precursor ions is obtained. Each candidate peptide in the list is then used to generate a spectrum of respective theoretical product ions series. By comparing them to the MS/MS data from the experiment, the matched peptides and the corresponding proteins are ranked and scored, from the most probable to the least.

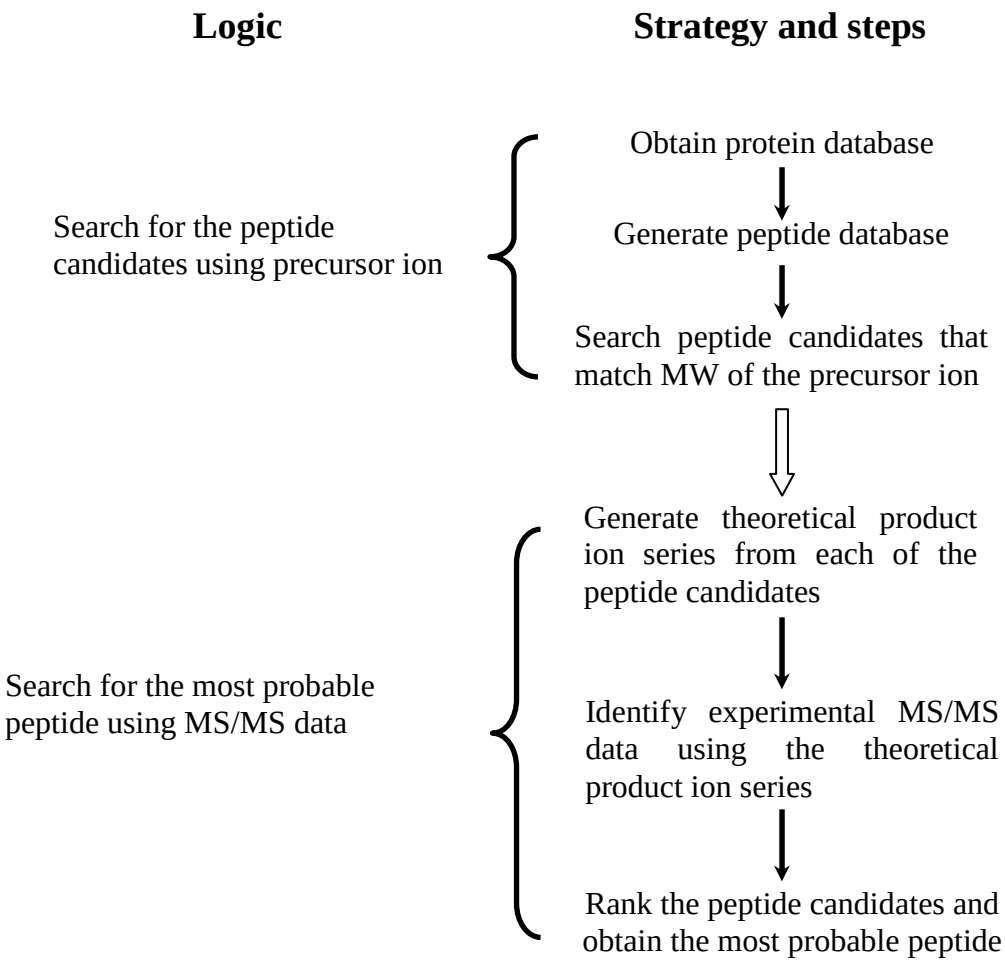


Figure 4.1 General strategies for peptide and protein identification

4.3.2 Ranking criteria

Since different ranking and scoring strategies are implemented, different peptides and proteins may be identified; particularly, when a whole-cell proteome is to be analyzed. Instead of using probability-based techniques for MS/MS interpretation, we used a heuristic approach for identifying protein in a mixture. By comparing searched results obtained from other tools and ours, more confident conclusions may be drawn.

The criteria to rank peptide candidates in this proposed approach included: 1) counting the total numbers of identified MS/MS data (T); 2) tallying the total number of identified b- and y- type ions (T_{b+y}); 3) calculating the percentage of T_{b+y} in T if the identified ions contain other ions (i.e., b-H_2O, b-NH_3, y-H_2O, and y-NH_3); and 4) locating the contiguity of identified ions. In practice, the first step of the heuristic approach is to select the peptides with T_{b+y} greater than 50% of T, because b- and y- type ions are generally considered as the major product ions after fragmentation. The second step is to rank these selected peptides in descending order according to T. Several ranking steps are involved in the last process: 1) peptides are ranked according to T if none of them have the same T; 2) when peptides have the same T but different T_{b+y}, they will be ranked in descending order according to T_{b+y}; 3) when peptides have both the same T and T_{b+y}, a ranking is made based on the contiguity of identified ions. As an example, a peptide having identified b_5 and b_6 ions will be ranked higher than another peptide having b_5 and b_8 ions; and 4) if peptides have the same T, T_{b+y} and the contiguity of b-

or y-type ions, the corresponding precursor ion is regarded as un-identified and discarded from the subsequent report.

4.4 Implementation

To validate the proposed approach, the MS/MS data of a known peptide sequence and an unknown peptide were used. The known peptide sequence reported by Peng *et al*. (2003) was adopted, and the mass information for both precursor ion and the corresponding product ions series were collected in Table 4.1 as known peptide sequence data. The original MS/MS data (000.30.30.2.dta) for *S. cerevisiae* was downloaded from http://bioinformatics.icmb.utexas.edu/OPD (Prince *et al*., 2004). The required mass information of precursor and product ions were compiled and shown in Table 4.1 as unknown peptide sequence data.

Table 4.1 Experimental MS spectral data* extracted from Peng *et al.* (2003) and Prince *et al.* (2004)

Known peptide	Precursor ion (m/z, z)	1010.7, +2
	Product ions series (m/z)	538.54, 609.62, 722.78, 779.83, 850.91, 899.07, 964.07, 986.15, 1035.15, 1057.22, 1122.22, 1170.38, 1241.46, 1298.51, 1411.67
Unknown peptide	Precursor ion (m/z, z)	1578.75, +2
	Product ions series (m/z)	242.3, 314.1, 379.3, 389.1, 393.1, 413.3, 416.2, 433.3, 434, 445, 476.2, 496, 497.3, 500, 501.1, 502.9, 506.2, 511.2, 514.7, 528.9, 541, 557.3, 571.8, 576.2, 593.8, 597.1, 599, 600.9, 609.9, 615.5, 616.5, 628.8, 637, 637.7, 640.1, 640.9, 641.7, 645.5, 647.9, 654.8, 658.3, 663.5, 669.3, 671.6, 673, 673.8, 679.6, 687.4, 689.8, 691.1, 693.4, 698.9, 700.3, 703.2, 707, 708.3, 715.8, 718.7, 722.2, 725.2, 727.3, 730.2, 730.9, 736.1, 742.6, 743.6, 744.3, 748.7, 750.5, 751.9, 752.9, 753.7, 754.4, 760.9, 770.5, 771.4, 772.4, 773.4, 788.1, 789, 792.9, 822.4, 823, 827.8, 836, 853.4, 868.4, 871.7, 873.3, 876, 892.4, 893.1, 899.5, 940.2, 994.4, 999.3, 1001.3, 1021.4, 1023.8, 1025.2, 1058.1, 1098.6, 1114.8, 1185.4, 1186.2, 1207.3, 1209.5, 1215.7, 1245.8, 1256.4, 1257.5, 1258.2, 1259.1, 1344.3, 1345.1, 1427.2, 1480.7

*Precursor ion is shown as mass-to-charge ratio (m/z) followed by charge state. All m/z values of both precursor ion and product ion series are monoisotopic mass.

The protein database of *S. cerevisiae* (Swiss Prot 42.6) was retrieved to construct an *in silico* trypsin-digested peptide database. Additionally, the following two protein modifications were also taken into consideration during peptide database reconstruction: 1) all cysteine residues were treated to form Cys_CAM; and 2) maximum missed cleavages = 1. During the course of identification, peptide molecular mass tolerance was set as 1.0 Da and the MS/MS ion series tolerances were set as 0.8 Da.

Both sets of MS/MS data were also applied to several protein interpretation tools including Mascot, MS-Tag, and SEQUEST, such that the effectiveness of the proposed heuristic approach could be verified. The identified most probable peptide (ranked first for both known and unknown sequence) by various tools was compiled in Table 4.2

Table 4.2 Searched results among several protein identification tools

	Tools	Ions identified*	Search results
Known peptide sequence at m/z =1010.7	Mascot	15/15	HEAAEALGAIASPEVVDVLK
	MS-Tag	15/15	HEAAEALGAIASPEVVDVLK
	Ours	15/15	HEAAEALGAIASPEVVDVLK
	SEQUEST	15/15	HEAAEALGAIASPEVVDVLK
Unknown peptide sequence at m/z =789.88	Mascot	15/117	KLEDAEGQENAASSE
	MS-Tag	31/117	KLEDAEGQENAASSE
	Ours	37/117	KLEDAEGQENAASSE
	SEQUEST	15/117	KLEDAEGQENAASSE

* Number of identified experimental MS/MS spectral data / Number of experimental MS/MS spectral data

It can be seen that the searched results were in agreement even though different scoring strategies were implemented. For instance, SEQUEST uses re-constructed MS/MS spectra as the basis and counts the number of experimentally obtained product ions that match the re-constructed spectra, followed with statistical reasoning to rank possible peptides. In contrast, we used product ions obtained from the experiment as the reference along with simple rules defined in Section 4.3.2, such that intricate calculations may be avoided during the peptide and protein interpretation. Detailed searched results for both known and unknown sequence peptides are included in Supplementary Data 1 and Supplementary Data 2 of Appendix C, respectively.

4.5 Concluding remarks

Searching for the most probable peptide and protein is the final and crucial step during the course of protein identification. One method uses theoretical product ions series as the basis, searching through the experimental data. Another method uses the experimental MS/MS data as the basis, attempting to match the theoretical counterparts. It is not easy to differentiate which approach is superior to others since the MS/MS data itself is extremely complex, particularly for a whole-cell proteome. Different numbers of ions were identified as different tools were employed, implying that the discrepancy of the searched results would be enlarged. Thus, it is recommended that at least two or more protein interpretation tools should be applied to the same set of MS/MS data, such that a higher level of confidence on the identified proteins is obtained.

4.6 References

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

Clauser, K.R., Baker, P. and Burlingame, A.L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem*, **71**, 2871-2882.

Eng, J.K., McCormack, A.L. and Yates, J.R., III. (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.

Gygi, S.P. and Aebersold, R. (2000) Mass spectrometry and proteomics. *Curr Opin Chem Biol*, **4**, 489-494.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M. and Yates, J.R., III. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol*, **17**, 676-682.

Mann, M. and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem*, **66**, 4390-4399.

Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. and 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.

Peng, J. and Gygi, S.P. (2001) Proteomics: the move to mixtures. *J Mass Spectrom*, **36**, 1083-1091.

Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551-3567.

Prince, J.T., Carlson, M.W., Wang, R., Lu, P. and Marcotte, E.M. (2004) The need for a public proteomics repository. *Nat Biotechnol*, **22**, 471-472.

Qin, J., Fenyo, D., Zhao, Y., Hall, W.W., Chao, D.M., Wilson, C.J., Young, R.A. and Chait, B.T. (1997) A strategy for rapid, high-confidence protein identification. *Anal Chem*, **69**, 3995-4001.

Washburn, M.P., Wolters, D. and Yates, J.R., III. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, **19**, 242-247.

Chapter 5 Case studies: growth of *Saccharomyces cerevisiae* under low and high glucose conditions

This chapter describes the implementation of results deduced from our developed two-step strategy (Section 3.2) and two-pass approach (Chapter 4) by studying the protein profiling of *S. cerevisiae* grown under different glucose concentrations.

The contents of this chapter were subdivided into two major parts. Firstly, *S. cerevisiae* was cultivated in 4 different glucose concentrations (10, 100, 200, and 300 g glucose/l) and the changes of ethanol and glucose concentrations in these conditions were compared. Secondly, protein profiling of *S. cerevisiae* cells harvested under 10 and 300 g glucose/l was used in an attempt to interpret the lower ethanol yield under 300 g glucose/l condition. Detailed descriptions of these subjects are presented below.

5.1 Growth of *Saccharomyces cerevisiae* in a chemostat under high glucose conditions

The content in this section was published in *Biotechnology Letters,* **25**, 1151-1154 (2003).

5.1.1 Abstract

A chemostat apparatus was used to cultivate *Saccharomyces cerevisiae* under high glucose conditions (up to 300 g/l). The results support the view that higher glucose feed favours higher ethanol production regardless of the existence of osmotic stress. A low glucose utilization and yield coefficient provides an opportunity to improve continuous fermentation performance in the fuel alcohol industry. To reuse yeast cells and subsequently lower operating cost, an optimal glucose feeding concentration (between 100 and 200 g/l) exists.

Nomenclature*

$C_{i,o}$	Initial concentration of metabolite i (g/l)
$C_{i,ss}$	Steady-state concentration of metabolite i (g/l)
X	Biomass (g dry wt/l)
t_g	Generation time (h)
μ	Specific growth rate (1/h)

*subscript i represents either ethanol (p) or glucose (s) used in the text.

5.1.2 Introduction

The 1997 Kyoto Protocol has gathered international attention. Since fossil fuel-powered vehicles produce a large percentage of greenhouse gas emissions, there is a growing movement to produce fuels based on renewable resources that are more environmentally friendly than traditional petrochemicals. Ethanol, as a fuel additive (a 10% blend),

reduces the emission levels of CO (up to 30%), CO_2 (up to 10%), and smog-causing hydrocarbons (up to 7%).

Batch fermentation is the traditional practice in the fuel alcohol industry for ethanol production following dry milling of grain. Alcohol concentrations of 10 – 12% have been the norm. To increase the ethanol productivity and profits per batch, a higher sugar feed can be dosed. Thomas *et al*. (1993) reported that as much as 23.8% (v/v) ethanol can be made from 38 °P dextrinized starch (°P = grams dissolved solids measured as sucrose per 100 g of mash) in a laboratory batch fermenter with all substrates present at zero time using normal commercial active dry yeasts. A high glucose feed would impose a serious stress to *S. cerevisiae*; this stress would cause slow cell proliferation and a decline of cell viability (Thomas & Ingledew, 1992). In addition, substrate-accelerated death of cells (Teusink *et al*., 1998) is accentuated although this is not a problem as this industry does not normally reuse their yeasts following batch fermentation.

Batch fermentation features ease of operation and closer control of bacterial contamination. However, some pre-fermentation processes such as cleaning, sanitizing filling, and emptying all take time, representing a major loss of productivity. Due to the transient nature of batch fermentation, ethanol production also varies with time, making it difficult for process analyses. As an alternative, continuous fermentation maintains the process at steady state for any given period, and leads to constant production rates, making it easier for process optimization toward high ethanol yield. Continuous fermentation is also easy to control during steady state, thus reducing the downtime as

observed in a batch operation. More detailed comparisons between batch and continuous fermentation are found in Kelsall and Lyons (1999).

Currently (2003), ethanol production in North America is about 10 billion l/yr, and is expected grow to 14 billion l/yr or more by the end of 2005. Since the profit margin of the fuel alcohol industry is relatively low, techniques that can increase ethanol production without increasing investment are needed. In this study, *S. cerevisiae* was cultivated in a chemostat apparatus in an attempt to investigate the influence of high glucose feed and specific growth rate on ethanol yield and yeast response. Modifications and suggestions to increase ethanol productivity in a chemostat are provided.

5.1.3 Materials and methods

5.1.3.1 Yeast and culture conditions

Saccharomyces cerevisiae originally supplied by Alltech Co. (Nicholasville, KY) and held in pure culture at Dr. W. M. Ingledew's laboratory at the University of Saskatchewan, Canada, was used in this study. A chemically defined medium adapted from Narendranath *et al*. (2001) was used in this study. The medium contained either 10, 100, 200, or 300 g glucose/l as the sole carbon source; the $(NH_4)_2SO_4$ was fixed at 2.64 g/l to avoid the nitrogen growth-limiting effect (Thomas *et al*., 1996) and vitamins used were 1.5 times the concentrations used by Narendranath *et al*. (2001).

During the experiments, a multi-head peristaltic pump was used to deliver fresh medium and withdraw spent broth from a 2 liter fermenter (Model: Virtis Omni-Culture, Virtis Inc., NY) at a dilution rate of 0.12 h^{-1}. During these runs, working volume, temperature, and agitation rate were maintained at 1.0 l, 28°C and 100 rpm. Sterile air was flushed only to the headspace region of the fermenter at 0.2 l/min to allow the yeast to synthesize required unsaturated fatty acids and sterols while still maintaining an anaerobic environment for yeast cells for ethanol production (O'Connor-Cox & Ingledew, 1989).

5.1.3.2 Sample analysis

Once a steady state was reached where the specific growth rate equaled the dilution rate, a generation time was estimated from the dilution rate using the equation: t_g = ln 2/dilution rate. Generally, a time of 10 times the yeast generation time is sufficient for a microbial population to reach balanced growth (Gostomski *et al.*, 1994). After that, five consecutive samples were taken spaced one-generation time apart. Total numbers of yeast cells and cell viability were determined microscopically with the aid of methylene blue (Thomas & Ingledew, 1990). Biomass dry weight was determined by centrifuging 50 ml samples at 14,600 g (4°C) for 15 min, washing the cell pellet twice with cold water, and drying the pellet overnight at 65°C in a vacuum oven under a pressure of 70 kPa. Glucose and ethanol in the supernatant were measured using an ORH-801 column (Transgenomic Co., NE) on an HPLC (Model 1100 series, Agilent Technologies, CA) equipped with a refractive index detector (HP1047A). The column was eluted at 65°C with 5 mM H_2SO_4 at 0.3 ml/min.

5.1.4 Results and discussion

At each glucose dose, five samples (one generation-time apart) were withdrawn, analyzed, and the measured results were quantified and averaged (Table 5.1). The working volume of the fermenter used and flow rate were examined before and after each experiment to ensure that a constant specific growth rate (μ) has been maintained. Four specific growth rates were obtained as 0.123, 0.123, 0.128, and 0.130 h^{-1} (average = 0.126 h^{-1}; standard deviation = 0.004 h^{-1}) corresponding to glucose feeds at 10, 100, 200, and 300 g/l, respectively. The biomass under these four glucose doses were 0.729 ± 0.019, 0.905 ± 0.003, 0.708 ± 0.034 and 0.646 ± 0.018 g dry wt/l. A decline in biomass was observed as the glucose increased from 100 to 300 g/l. Such a trend might be attributed to the osmotic effect contributed by high glucose concentrations, resulting in slower proliferation of yeast cells. This trend is in agreement with the report of Thomas and Ingledew (1992). Those authors also found that the cell viability decreased as the sugar concentration increased during ethanol fermentation; however, a similar phenomenon was not observed in our current studies, where over 90% of yeast viability was recorded for all runs. One possible explanation for the discrepancy might be the viscosity effects resulting from different medium formulations (that is, defined versus complex) used in this work and that of Thomas and Ingledew (1992). The viscosity in defined media was relatively lower than that of complex ones, allowing the metabolic CO_2 to easily escape from the broth. Otherwise, accumulated CO_2 would consequently inhibit yeast growth (Thomas *et al*., 1994).

Table 5.1 Concentrations of residual glucose and ethanol under various glucose concentrations

Glucose concentration (g/l)	Residual glucose (g/l)	Ethanol (g/l)
10	0.52 ± 0.09	3.67 ± 0.25
100	36.28±2.09	14.18 ± 1.80
200	106.59 ± 1.53	24.48 ± 1.68
300	136.34 ± 7.47	40.03 ± 4.42

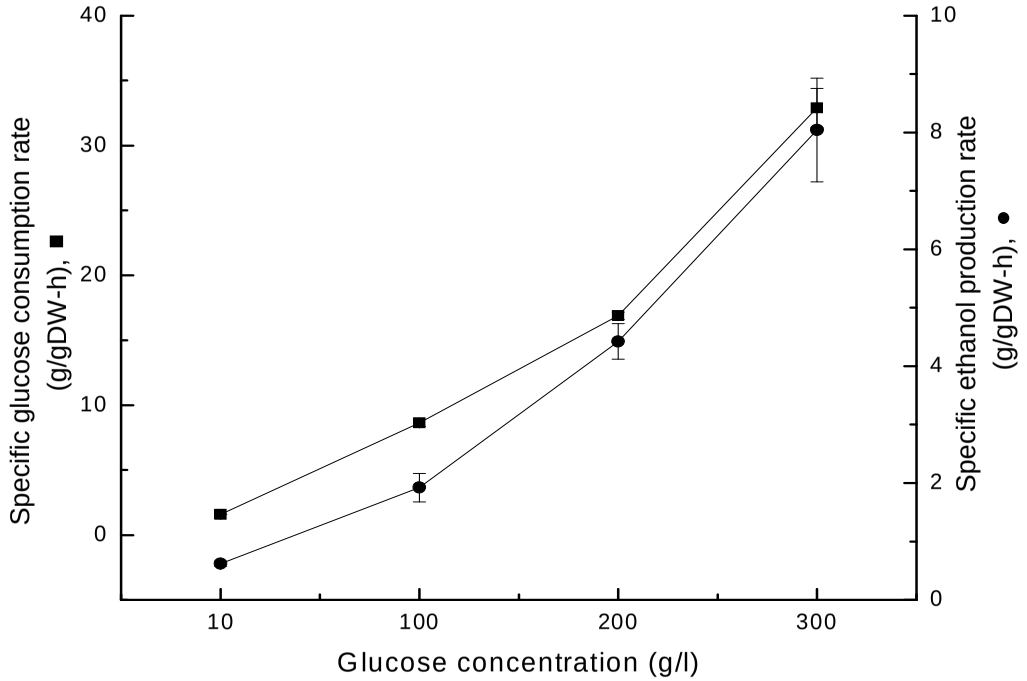


Figure 5.1 Specific consumption (SGCR) and/or production (SEPR) rates of glucose and ethanol by *S. cerevisiae* grown in a chemostat under various glucose concentrations

The specific consumption and/or production rates (defined as $\mu(C_{i,ss}–C_{i,o})/X$) and the yield coefficient (defined as $Y_{p/s} = (C_{p,ss}–C_{p,o})/(C_{s,o}–C_{s,ss})$) are commonly used to assess microbial performance. Both criteria reflect different aspects of meaning in evaluating a bioprocess operation (Lin *et al*., 2002). Figure 5.1 shows that a higher glucose dose results in a higher specific glucose consumption rate (SGCR) and a higher specific ethanol production rate (SEPR) under the same growth condition, indicating an increased metabolic flux through the glycolytic pathway leading to ethanol.

Yield coefficient changes with glucose feed are seen in Figure 5.2. Narendranath *et al*. (2001) reported that an $Y_{p/s}$ of 0.40 was attained when growing the same strain with 20 g glucose/l in a batch culture. Results obtained from our current investigation in a chemostat apparatus with 10 g glucose/l feed were compatible with their findings. This might indicate that a continuous fermentation can reach almost the same ethanol production as batch fermentation. However, Figure 5.2 also illustrates that a lower glucose feed correlated with a higher $Y_{p/s}$, and no significant increase of $Y_{p/s}$ was noticed as glucose feed increased over a threshold concentration (likely 100 g glucose/l under current investigation conditions). It seems that the carbon fraction channelling to ethanol synthesis was saturated when glucose concentration was over 100 g/l, and thus $Y_{p/s}$ was kept at 0.24 ± 0.03. This observation is similar to that reported by Bayrock and Ingledew (2001), who cultivated the same strain using a complex medium containing glucose varying from 152 to 312 g/l in a multi-stage continuous fermentation. In their report, a constant $Y_{p/s}$ of 0.38 was obtained irrespective of glucose feed.

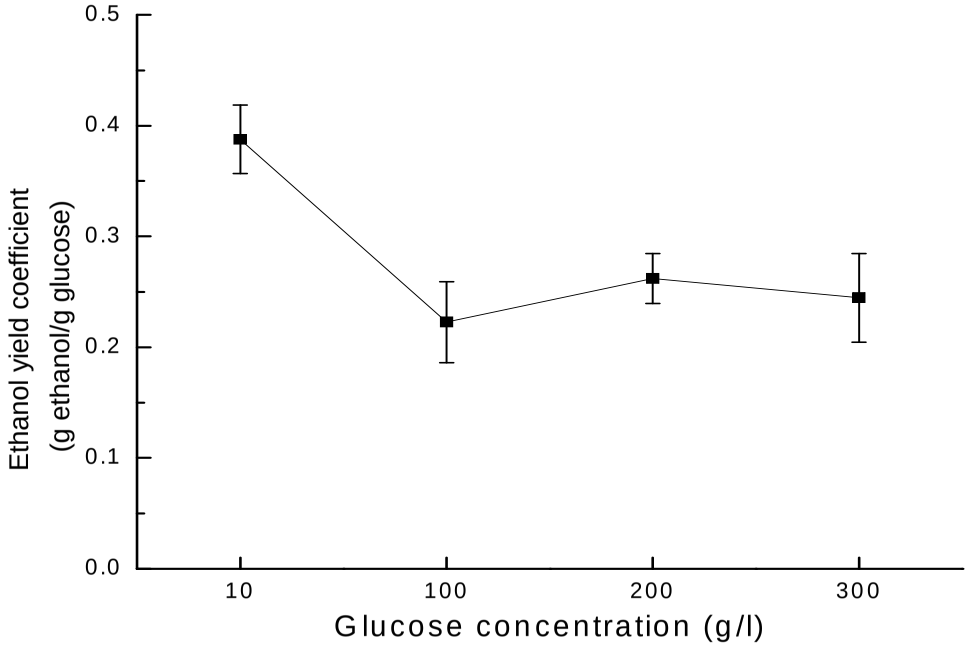


Figure 5.2 Ethanol production yield coefficient of *S. cerevisiae* grown in a chemostat under various glucose concentrations

Comparatively, $Y_{p/s}$ calculated from the current study at a glucose feed ≥ 100 g/l was lower than that obtained by Bayrock and Ingledew (2001). The reason might be attributed to the elevated and combined osmotic effects due to the presence of high glucose concentration and the accumulation of ethanol in the fermentation broth. *S. cerevisiae* grown under such conditions will alter its metabolic regulation to adapt to the harsh environment for survival. For instance, yeasts might (1) synthesize and excrete metabolites such as glycerol and trehalose, to protect cellular integrity (Mansure *et al*., 1997); and (2) generate and/or regulate the energy and reducing power required for growth to avoid substrate-accelerated death (Teusink *et al*., 1998). In a complex medium, osmoprotectants might already be present, saving glucose flux and channelling glucose toward ethanol synthesis. Comparatively, these osmotic regulating substances might not exist in a defined medium, such that a fraction of carbon from glucose would be utilized for synthesizing protecting compound(s) to overcome stressful conditions, ultimately resulting in low $Y_{p/s}$. In this study, we have also observed that concentrations of proline, glycerol, and trehalose (common osmoprotection chemicals) increased concurrently with glucose feed.

5.1.5 Concluding remarks

Driven by environmental concerns, the demand for fuel alcohol is increasing. To meet the demand, a fermentation process featuring a higher ethanol production and economic feasibility is preferable. A continuous operation becomes an apparent choice. Single-stage continuous fermentation can only utilize a fraction of the carbon source (e.g., in

this study, about 54% of glucose was utilized for ethanol synthesis, resulting in a relatively lower $Y_{p/s}$ than a batch operation). In contrast, a multi-stage continuous fermentation converts more sugar to ethanol and maintains a relatively high apparent $Y_{p/s}$ (Bayrock & Ingledew, 2001). Cultivating yeast cells at higher dilution rates would result in higher SCGR and SEPR, but lower ethanol yields. On the other hands, lower dilution rates favor higher ethanol yields, but low dilution rates also prolong operations prior to attainment of steady-state conditions.

Since no appreciable change in $Y_{p/s}$ was observed when cultivating *S. cerevisiae* under higher glucose conditions, it could then be extrapolated that a higher ethanol production with a nearly zero glucose discharge could be obtained in a multi-stage continuous fermentation operation. Until now, very limited literature information on this topic is available (Bayrock & Ingledew, 2001; Lin *et al*., 2002). Further study should focus on the selection of dilution rates, the number of fermentation stages, the optimal glucose concentration, the recycle ratio of yeast cells, the maximum alcohol concentrations which can be obtained, and the interactive effects of these parameters during high glucose continuous fermentation.

5.1.6 Acknowledgments

Authors are grateful to the Natural Sciences and Engineering Council of Canada for financial support. Particular thanks are due to Prof. W.M. Ingledew at the University of Saskatchewan for suggestions and critical reading of the manuscript.

5.1.7 References

Bayrock, D.P. and Ingledew, W.M. (2001) Application of multistage continuous fermentation for production of fuel alcohol by very-high-gravity fermentation technology. *J Ind Microbiol Biotechnol*, **27**, 87-93.

Kelsall, D.R. and Lyons, T.P. (1999) Management of fermentations in the production of alcohol: moving toward 23% ethanol. In: Jacques, K.A., Lyons, T.P. and Kelsall, D.R. (eds). *The Alcohol Textbook: A Reference for the Beverage, Fuel and Industrial Alcohol Industries*, 3rd edn. UK: Nottingham University Press.

Gostomski, P., Muhlemann, M., Lin, Y.-H., Mormino, R. and Bungay, H.R. (1994) Auxostats for continuous culture research. *J Biotechnol,* **37,** 167–177.

Lin, Y.-H., Bayrock, D.P. and Ingledew, W.M. (2002) Evaluation of *Saccharomyces cerevisiae* grown in a multi-stage chemostat environment under increasing levels of glucose. *Biotechnol Lett*, **24**, 449-453.

Mansure, J.J., Souza, R.C. and Panek, A.D. (1997) Trehalose metabolism in *Saccharomyces cerevisiae* during alcoholic fermentation. *Biotechnol Lett*, **19**, 1201-1203.

Narendranath, N.V., Thomas, K.C. and Ingledew, W.M. (2001) Effects of acetic acid and lactic acid on the growth of *Saccharomyces cerevisiae* in a minimal medium. *J Ind Microbiol Biotechnol*, **26**, 171-177.

O'Connor-Cox, E.S.C. and Ingledew, W.M. (1989) Effect of the timing of oxygenation on very high gravity brewing fermentations. *J. Am. Soc. Brew. Chem.* 48: 26-32.

Teusink, B., Walsh, M.C., van Dam, K. and Westerhoff, H.V. (1998) The danger of metabolic pathways with turbo design. *Trends Biochem Sci*, **23**, 162-169.

Thomas, K.C. and Ingledew, W.M. (1990) Fuel alcohol production: effects of free amino nitrogen on fermentation of very-high-gravity wheat mashes. *Appl Environ Microbiol*, **56**, 2046-2050.

Thomas, K.C. and Ingledew, W.M. (1992) Production of 21% (v/v) ethanol by fermentation of very high gravity (VHG) wheat mashes. *J Ind Microbiol*, **10**, 61-68.

Thomas, K.C., Hynes, S.H., Jones, A.M. and Ingledew, W.M. (1993) Production of fuel alcohol from wheat by VHG technology: effect of sugar concentration and fermentation temperature. *Appl Biochem Biotechnol*, **43**, 211-226.

Thomas, K.C., Hynes, S.H. and Ingledew, W.M. (1994) Effects of particulate materials and osmoprotectants on very-high-gravity ethanolic fermentation by *Saccharomyces cerevisiae*. *Appl Environ Microbial*, **60,** 1519-1524.

Thomas, K.C., Hynes, S.H., Jones, A,M, and Ingledew, W.M. (1996) Effect of nitrogen limitation on synthesis of enzymes in *Saccharomyces cerevisiae* during fermentation of high concentration of carbohydrates. *Biotechnol Lett*, **18**, 1165-1168.

5.1.8 Additional experimental details

Yeast and media

The medium contained 10, 100, 200, or 300 g glucose/l as the sole carbon source; the $(NH_4)_2SO_4$ was fixed at 2.64 g/l. The final concentrations of other ingredients in the medium were, in millimoles per liter: K_2HPO_4, 0.86; KH_2PO_4, 6.83; $MgSO_4$, 2.03; NaCl, 2.05; in micromoles per liter: H_3BO_3, 24; $MnSO_4$ 20; Na_2MoO_4, 1.5; $CuSO_4$, 10; $CoCl_2$, 1.5; $ZnSO_4$, 100; KI, 1.8; $FeCl_3$, 100; $CaCl_2$, 82; and in micrograms per liter: biotin, 300; calcium pantothenate, 3,000; folic acid, 30; myoinositol, 15,000; niacin, 600; pyridoxine HCl, 600; riboflavin, 300; and thiamine HCI, 300. The vitamin solution was prepared as a 1,000-fold concentrated stock and kept frozen at -20°C. When needed, an aliquot was thawed and filter-sterilized (0.2-µm membrane filter). The rest of the components were weighed and subdivided into three parts for making solution: 1) glucose (total volume 10 l); 2) K_2HPO_4, KH_2PO_4, $MnSO_4$, Na_2MoO_4, KI, $CuSO_4$, H_3BO_3, $CoCl_2$, and $ZnSO_4$ (total volume 2 l); and 3) NH_4SO_4, $MgSO_4$, NaCl, $FeCl_3$, $CaCl_2$, and $MgSO_4$ (total volume 8 l). All parts of medium were autoclaved at 121°C for 50 min. After sterilization, all cooled parts and vitamin stock were aseptically combined to form the final medium. All microbiological medium ingredients were purchased from VWR Inc., ON.

Fermentation system and growth conditions

A typical chemostat fermentation technique with few modifications was used through this experiment (Figure 5.3). During the experiments, a multi-head peristaltic pump

(Model 7520-25, Cole-Parmer Instruments Co., IL) was used to deliver fresh medium and withdraw spent broth from a 2 l fermenter (Model: Virtis Omni-Culture, Virtis Inc., NY) at a dilution rate of 0.12 h^{-1}. During these runs, working volume, temperature, and agitation rate were maintained at 1.0 l, 28°C and 100 rpm. Sterile air was flushed only to the headspace region of the fermenter at 0.2 l/min to allow the yeast to synthesize required unsaturated fatty acids and sterols while still maintaining an anaerobic environment for yeast cells for ethanol production (O'Connor-Cox and Ingledew, 1990). In order to prevent its loss during the course of fermentation, a pre-sterilized condenser circulating with chilled water at 4°C was used and installed at the exhaust gas line just before the air filter (pore size = 0.2 µm).

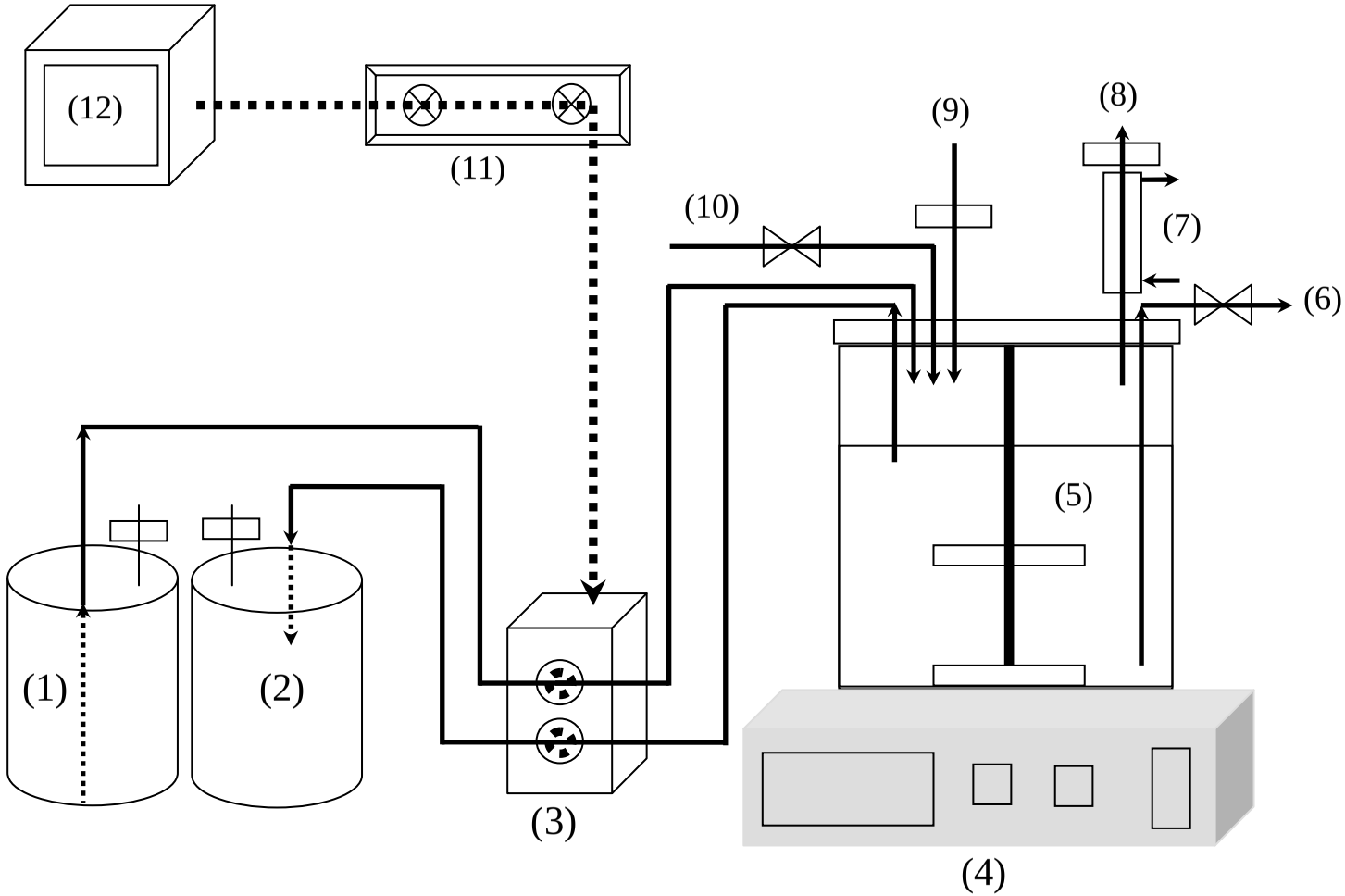


Figure 5.3 Schematics of chemostat fermentation system (1) Medium reservoir; (2) Spent broth reservoir; (3) Multi-head pump; (4) Fermentation station; (5) Fermenter; (6) Sampling line; (7) Condenser; (8) Air filter; (9) Air inlet; (10) Inoculation inlet; (11) Data acquisition board; (12) Computer control system.

5.2 A proteomic study of *Saccharomyces cerevisiae* grown under high specific gravity conditions

Part of the contents in this section was presented at the 54th Canadian Chemical Engineering Conference at Calgary (Canada), October 3-6, 2004. This section applied the two-pass approach (Chapter 4) and Mascot to identify protein profiling of *S. cerevisiae* grown at 10 and 300 g glucose/l; and the proteins identified with high confidence were then deduced using the developed two-pass strategy (Section 3.2). Finally the proteins distributed in central metabolic pathways were compared to interpret experimental observations. This section is prepared in a manuscript format for possible journal publication.

5.2.1 Abstract

Multi-dimensional protein identification technology (MudPIT) was implemented to investigate the protein expression profile of *S. cerevisiae* grown under two different specific growth conditions (i.e., 10 and 300 g glucose/l). The experimental results show that the proteins associated with the pentose phosphate (PP) pathway and the anaplerotic pathway (PYR + CO_2 → OAA) under a 300 g glucose/l condition are identified compared with those under 10 g glucose/l conditions, indicating that more metabolic flux was diverted into the PP pathway and the TCA "cycle" in order to survive osmotic stresses resulting from the high specific gravity fermentation. These observations may partially be used to explain why the relative yield of ethanol from glucose is comparatively low under a high glucose condition although the total ethanol production

is high under 300 g glucose/l condition as opposed to that under the 10 g glucose/l condition.

Abbreviation of metabolites

ACCOA	Acetyl-coenzyme A
AKG	α-Ketoglutarate
CO_2	Carbon dioxide
E4P	Erythrose-4-phosphate
F-1,6-P	Fructose-1,6-bisphosphate
F6P	Fructose-6-phosphate
G6P	Glucose-6-phosphate
GAP	Glyceraldehyde-3-phosphate
GLC	Glucose
ISOCIT	Isocitrate
MAL	Malate
OAA	Oxaloacetate
6-PGCLAC	6-Phosphogluconolactone
6-PGC	6-Phosphogluconate
PEP	Phosphoenolpyruvate
PYR	Pyruvate
R5P	Ribose-5-phosphate
RU5P	Ribulose-5-phosphate
S7P	Sedoheptulose-7-phosphate
SUC	Succinate
X5P	Xylulose-5-phosphate

List of Genes

Gene name	Description
ENO1	Enolase 1
GLK1	Glucokinase
GND2	Phosphogluconate dehydrogenase
GPM2	Phosphoglycerate mutase
HXK2	Hexokinase isoenzyme 2
HXT1	Hexose transporter
HXT3	Hexose transporter
HXT4	Hexose transporter
HXT5	Hexose transporter
PYK2	Pyruvate kinase
RPE1	D-ribulose-5-phosphate 3-epimerase
RPI1	Small GTPase regulatory/interacting protein
STL1	Sugar transporter-like protein
TDH1	Glyceraldehyde-3-phosphate dehydrogenase 1
YBR241C	Putative hexose transporter

5.2.2 Introduction

High gravity and/or very-high-gravity fermentation have become effective methods to produce ethanol to meet world demand, because a higher ethanol production is always obtained under higher sugar concentration (Bayrock and Ingledew, 2001; Thomas *et al*., 1993). The method is now well used in the ethanol industry. A high glucose feed, however, could impose a serious stress to *Saccharomyces cerevisiae*; this stress would cause slow cell proliferation and a decline in cell viability (Thomas and Ingledew, 1992). In addition, substrate-accelerated death of cells (Teusink *et al*., 1998) is accentuated. To survive in a stressful environment, *S. cerevisiae* has to modify its

metabolism, particularly the central metabolic routes (e.g., glycolysis pathway, pentose phosphate pathway, and tricarboxylic acid 'cycle') by triggering signal transduction systems and activating transcription and translation processes (Estruch, 2000; Ruis and Schuller, 1995). The survival of a yeast cell depends on its ability to quickly adapt to the changing environment. This ability is especially important for microbial industries, in which microorganisms are frequently subjected to various stress situations; their metabolic adaptations directly impact the production and economical profits.

Corresponding to the change of an organism's metabolism, the protein profile would be expected to vary significantly under stressful situations. Proteomics is the study of all expressed proteins of an organism grown under a given condition. The information obtained by comparing proteins under various conditions can help in dissecting the physiological states or phenotypic characteristics of an organism. In this study, *S. cerevisiae* was cultivated in a chemostat apparatus in an attempt to investigate the influence of high glucose feed on ethanol yield and on the yeast's response to a stress condition.

5.2.3 Materials and methods

Detailed information such as chemostat fermentation and sample treatment were described previously (Zhao and Lin, 2003). The free amino acids in the fermentation broth were reacted with Waters AccQ•fluor reagent, and then separated by an HPLC and quantified by measuring the absorbance using a scanning fluorescent detector (Model: Waters 474, Waters, Milford, MA, excitation = 250 nm, emission = 395 nm) as

described by the manufacturer (Waters, Milford, MA). The protein profiling was analyzed using multiple-dimensional HPLC separation coupled with a tandem mass spectrometry with the aid of protein identification tools and advanced identification confidence analysis strategy (Please see Additional Experimental Details).

5.2.4 Results and discussion

Protein profiling (Table 5.2) of *S. cerevisiae* grown at 10 and 300 g glucose/l conditions was used to elucidate why a relative low ethanol yield was observed under a high specific gravity condition. The proteins pertinent to central metabolism pathways, including the glycolytic pathway, pentose phosphate (PP) pathway, and tricarboxylic acid (TCA) cycle (Figure 5.4), were assessed and compared.

Table 5.2 Identified enzymes in the central metabolic pathway of *S. cerevisiae* grown at various glucose concentrations

Pathways	Enzyme	Glucose concentrations (g glucose/l)	
		10	300
Glycolysis	Hexokinase	√	√
	Phosphohexose isomerase		√
	Phosphofructokinase		√
	Aldolase	√	√
	Triose phosphate isomerase	√	√
	Phosphoglyceraldehyde dehydrogenase	√	√
	3-phosphoglycerate kinase	√	√
	Phosphoglyceromutase	√	√
	Enolase	√	√
	Pyruvate Kinase	√	√
Pentose phosphate	Glucose-6-phosphate dehydrogenase		√
	6-phosphogluconate dehydrogenase	√	√
	Transketolase		√
TCA ‘cycle’	Citrate synthase		√
	Pyruvate carboxylase		√

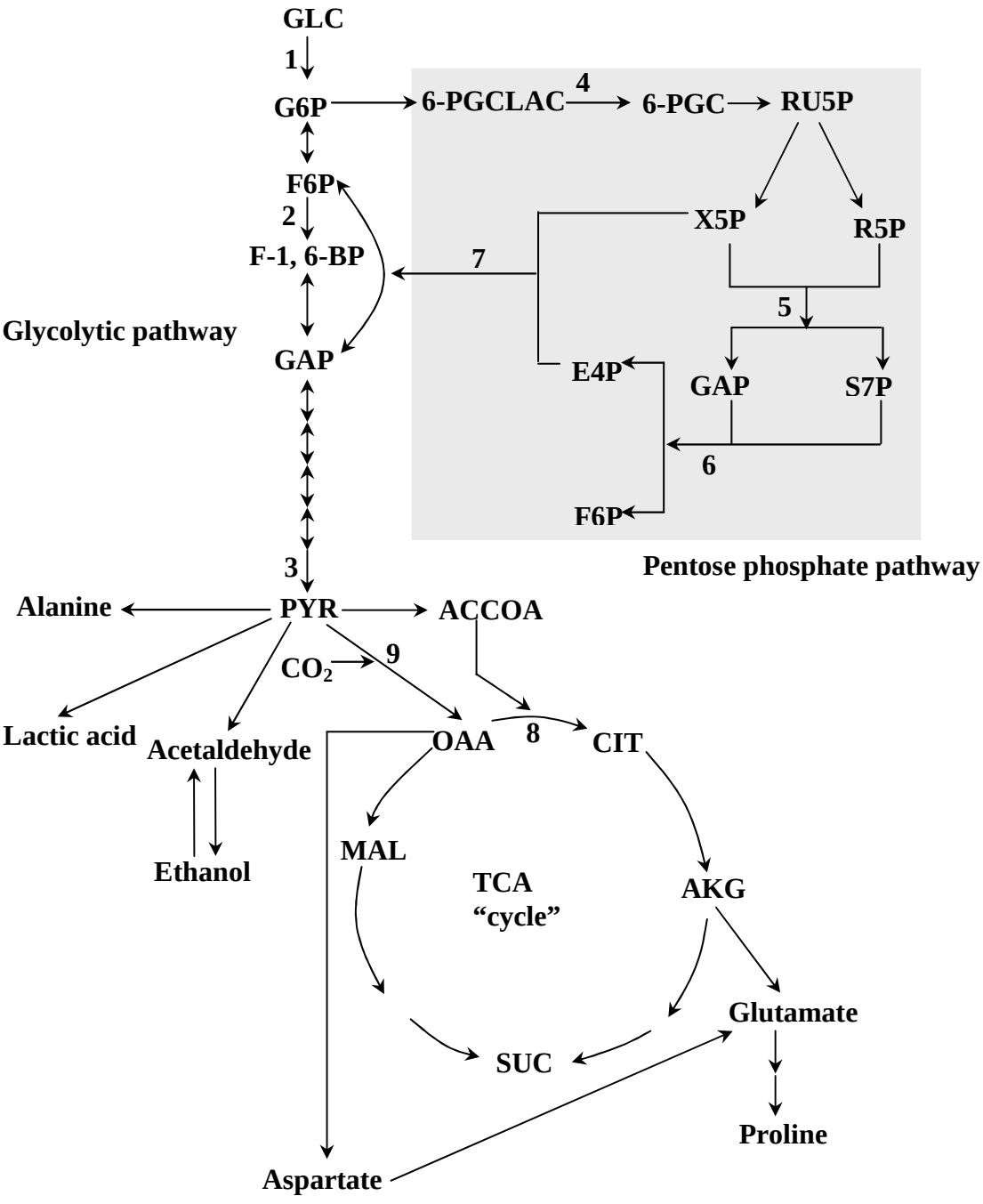


Figure 5.4 Schematics of central metabolic pathways used in *S. cerevisiae*
1, Hexokinase (HXK); 2, Phosphofructokinase (PFK); 3, Pyruvate kinase (PYK); 4, 6-phosphogluconate dehydrogenase (PGDH); 5, Transketolase (TKL); 6, Transaldolase (TAL); 7, Transketolase (TKL); 8, Citrate synthase (CIT1); 9, Pyruvate carboxylase (PYC). Named intermediates in the pathway are found on page 110 (Abbreviation of metabolites).

The glycolytic scheme pathway is the most common dissimilatory pathway and is found in nearly all living organisms. This pathway uses glucose as the starting substrate to carry out a sequence of ten biochemical reactions producing carbon skeletons and a relatively small amount of energy (ATP) and reducing power (NADH). The energy and reducing power, as well as the intermediates produced from glucose degradation are essential factors for a cell growth; they are used by the cell to synthesize the building block molecules (e.g., amino acids, enzymes and structural proteins, polysaccharides, and lipids) needed for cell propagation.

Generally speaking, there are ten enzymes involved in the glycolytic pathway, among which, hexokinase (HXK, EC:2.7.1.1) and/or glucokinase (GLK, EC:2.7.1.2), phosphofructokinase (PFK, EC: 2.7.1.11) and pyruvate kinase (PYK, EC: 2.7.1.40) play dominant roles in modulating the reaction rate of this pathway. In this study, HXK and GLK were identified when *S. cerevisiae* was grown under 10 g glucose/l, whereas only GLK was identified under 300 g glucose/l. The reason for this observation is that HXK is inhibited by its product, glucose-6-phosphate (G6P), while GLK is not subject to product inhibition by G6P. Under a higher glucose feeding condition, the specific consumption rate of glucose was increased (Zhao and Lin, 2003), meaning that the cell consumed more glucose and correspondingly more flux to G6P synthesis would be produced, resulting in the inhibition of HXK. Erasmus *et al*. (2003) used cDNA microarray technology to probe how the gene profiling changed when yeast cells were grown under 220 and 400 g sugar (equimolar amounts of glucose and fructose)/l conditions. Their results showed that the majority of hexose transporter-related genes including *HXT1*, *HXT5*, *STL1*, and YBR214C were up-regulated at higher sugar

conditions, while *HXT3* and *HXT4* were down-regulated. Nevertheless, the overall contribution of hexose transport genes was up-regulated, which is in agreement with our previous reported results (Zhao and Lin, 2003). Additionally, the *GLK1* encoding GLK was found to be up-regulated and *HXK2* encoding HXK was found to be down-regulated as the sugar concentration increased, which supported our experimental observation on protein profiling.

Microarray data published by Erasmus *et al*. (2003) showed that there was no significant change of the gene controlling PFK expression when yeast cells were grown under 220 to 400 g sugar/l conditions. In our experiment, PFK was not identified under eighter of the two glucose concentrations conditions, meaning the amount of PFK was very low and did not change significantly under these conditions. PYK was identified under the two glucose conditions in our work. Although our protein results could not provide a quantitative comparison, the microarray data (Erasmus *et al*., 2003) showed that PYK, encoded by *PYK2*, was up-regulated under 400 g sugar/l, indicating that the higher concentration of sugar helps yeast cells expressing PYK.

Four other enzymes belonging to the glycolysis pathway were identified for *S. cerevisiae* grown under the above two glucose concentration conditions. They are aldolase (FBA), 3-phosphoglycerate kinase (PGK), phosphoglyceromutase (PGM), and enolase1 (ENO1). The microarray data showed that the PGM encoded by *GPM2*, and the ENO1 encoded by *ENO1* were up-regulated while the other two enzymes showed no changes (Erasmus *et al*., 2003). Phosphoglyceraldehyde dehydrogenase (PGADH) was identified only at 300 g glucose/l, indicating that the corresponding gene, *TDH1*, encoding this

enzyme was up-regulated under higher sugar concentration. Comparing the level of *TDH1* under 220 g sugar/l to that of 400 g sugar/l conditions, TDH1 was up-regulated under the higher sugar concentration (Erasmus *et al*., 2003).

The pentose phosphate (PP) pathway is a pathway that converts 6 carbon molecules of glucose to 5 carbon sugars and other carbon skeletons ranging from 1 carbon to 7 carbons in size. It also generates reducing power in the form of NADPH. The primary functions of this pathway are: (1) to generate reducing power (NADPH), for reductive biosynthesis reactions; (2) to provide the cell with ribose-5-phosphate (R5P), a building block of ATP, COA, NAD^+, FAD, RNA, and DNA; and (3) to generate a broad spectrum of carbon intermediates for amino acid and protein synthesis. The PP pathway has both an oxidative and a non-oxidative arm. The oxidation steps, from G6P to ribulose-5-phosphate (RU5P), occur at the beginning of the pathway and are the reactions that generate NADPH. The non-oxidative reactions, started from RU5P, are primarily designed to generate R5P, fructose-6-phosphate (F6P) and glyceraldehyde-3-phosphate (GAP).

In this work, only 6-phosphogluconate dehydrogenase (PGDH, EC:1.1.1.44) was identified for the oxidative steps under 10 and 300 g glucose/l conditions. The microarray data showed that *GND2*, encoding PGDH, was up-regulated under higher sugar concentration (Erasmus *et al*., 2003), indicating more nutrient flux was shunt into the PP pathway as glucose feeding concentration was increased. The enzyme transaldolase (TAL, EC:2.2.1.2) was also identified in this study under 10 and 300 g glucose/l conditions, while the enzyme transketolase (TKL, EC:2.2.1.1) was identified

only under the 300 g glucose/l condition. These results were also supported by the findings of Erasmus *et al*. (2003).

Depending on the need of a yeast cell for R5P, NADPH, and ATP, the PP pathway can operate in various modes to fulfill different cellular demands. Generally, the PP pathway can operate in three major routes: (1) the generated RU5P may be converted to R5P, a precursor for synthesis of nucleotides and nucleic acids; (2) the generated GAP and F6P may be converted to glucose-6-phosphate and re-enter the PP pathway, maximizing the formation of NADPH; and (3) the generated GAP and F6P may enter glycolysis to produce pyruvate and ATP. We postulate that Routes 2 and 3 play more significant roles than Route 1. This postulation is based on the observation that the genes *RPE1* encoding phosphopentose epimerase and *RKI1* encoding R5P isomerase were shown by Erasmus *et al*. (2003) to be down-regulated when yeast was grown under higher sugar concentration (400 g sugar/l), indicating that R5P required under high sugar condition was decreased and more RU5P was thus accumulated. Secondly, it is noted that one manner of yeast's adaptation to osmotic stress condition is to modify the membrane permeability and integrity by changing the amount and composition of saturated and unsaturated fatty acids in the membrane. The biosynthesis of fatty acids requires a large amount of NADPH. Therefore, recycling some six carbon sugar (F6P) into the PP pathway to maximize NADPH production is mandatory for cells grown under higher glucose concentration, a condition that normally results in a higher osmotic stress condition for the cell. Thirdly, ATP is required for a cell to maintain its life and replication; therefore, part of F6P and most GAP should enter the glycolytic pathway for ATP generation. This route (Route 3) is especially important for yeast grown under high

sugar concentrations, because it prevents (at least limits) the accumulation of fructose-1,6-biphosphate (F-1,6-BP). The accumulation of F-1,6-BP could use up the cell's phosphate pool and may lead to cell death, or at least interfere with growth (Blomberg, 2000; Teusink *et al*., 1998).

Theoretically, oxidation of three moles of G6P in the PP pathway generates three moles of CO_2 and three moles of RU5P or R5P; these R5P molecules are then converted back into two moles of F6P and one mole of GAP. Therefore, the total carbon flux entering into the glycolytic pathway after the PP pathway in the form of F6P and GAP is less than the original flux in the form of G6P. It is agreed that not all F6P or GAP re-enter into glycolytic pathway under high osmotic stress conditions, part is recycled into the PP pathway, making more carbon loss in the form of CO_2. Additionally, the accumulation of RU5P under osmotic stress conditions contributes to additional loss of carbon from the glycolytic pathway. Therefore, the real carbon flux remaining in the glycolytic pathway is relatively less (using one carbon as basis) under higher glucose concentrations than that under low glucose concentrations. This leads to a lower ethanol production yield (g ethanol produced/g glucose consumed) even though the total production (ethanol amount) was higher under 300 g glucose/l (Zhao and Lin, 2003).

The TCA 'cycle' is not a real cycle when *S. cerevisiae* is grown anaerobically (Gancedo and Serrano, 1989; Lin *et al*., 2002). It is split into two directions: the first path is from oxaloacetate (OAA) to succinate (SUC) via citrate (CIT) and the second path is from OAA to SUC via malate (MAL). While not producing significant ATP, this 'cycle' provides key intermediates for building the cellular components. For example, α-

ketoglutarate (AKG) is a precursor of glutamate and then glutamine, arginine and proline, whereas OAA is a precursor of aspartate. In our work, two enzymes were found at 300 g glucose/l condition, they were citrate synthase (CIT1, EC:2.3.3.1) and pyruvate carboxylase (PYC, EC:6.4.1.1).

CIT1 governs the first reaction in the pathway from OAA to SUC via CIT. Since no more enzymes were identified in our work, nor any change was found in the microarray data published by Erasmus *et al.* (2003), we postulate that the major function of this section of the TCA 'cycle' is to produce metabolic intermediates made by the yeast cells to survive osmotic stress. It is noted that AKG, an intermediate in this path, is the precursor of glutamate, which is then, in part, used to synthesize glutamine, arginine, and proline. These amino acid units are key building blocks of proteins, which are essential for cell proliferation. Proline, a key component for the cell's osmotic-adaptation process, is also known to function as a compatible solute to help the yeast cell to counteract the immediate outflow of water from the cell under osmotic conditions. As an osmoprotectant, proline can be synthesized in the cell or provided extracellularly (Thomas *et al*., 1994). Hence, we infer that the major role of the first section of the TCA 'cycle' is to provide AKG for the synthesis of proline. In our work, the proline concentration in the fermentation broth was found to be 12-fold higher in 300 g glucose/l condition than that of 10 g glucose/l condition (Table 5.3).

Table 5.3 Concentrations of alanine and proline under various glucose concentrations

	10g/l	300g/l
Alanine (μM)	4.35 ± 0.29	30.42 ± 1.58
Proline (μM)	7.86 ± 2.15	107.31 ± 6.18

PYC plays the role of fixing CO_2 using PYR to form OAA, a substrate involved in both the first and second direction of the TCA 'cycle'. OAA is also a precursor of aspartate. Since no enzymes relevant to the second path were identified, we infer that there are two trends the generated OAA may pass through. The first trend is to produce CIT, governed by CIT1, which was described as above. The second trend is to synthesize aspartate, which is then used to build proteins or be transformed to produce glutamate, and then to make proline, the vital osmoprotectant compound.

5.2.5 Concluding remarks

Intermediates, energy (ATP) and reducing power (NADH, NADPH and $FADH_2$) are three important elements for cell growth. Intermediates produced from the central pathways are used as precursors to synthesize macromolecules such as lipids and nucleic acids, which then served as building blocks for the daughter cells. The availability of energy, reducing power, and the balance between the two is essential to guarantee a cell to grow well.

Under stress conditions, *S. cerevisiae* adjusts itself to adapt to the new environment; however, such an adaptation is mainly to sustain the cell propagation not an adjustment to overproduce ethanol. Our results showed that *S. cerevisiae* adapts to high osmotic stress condition through (1) the high activation of the PP pathway to provide more reducing power source (NADPH) for the synthesis of macromolecules that constitute cells; and (2) the use of the two directions of the TCA 'cycle' to provide intermediates to synthesize proline to counteract osmotic stress exerted by high glucose feed.

The data here cannot give a quantitative comparation of the protein profiling in the central metabolic pathway. Plus the microarray values sometimes cannot represent the real physiological adaptation process. Therefore most of our conclusions need to be further validated by using other techniques such as 2D-PAGE for protein quantification or enzyme assay for activity analysis.

5.2.6 References

Bayrock, D.P. and Ingledew, W.M. (2001) Application of multistage continuous fermentation for production of fuel alcohol by very-high-gravity fermentation technology. *J Ind Microbiol Biotechnol*, **27**, 87-93.

Blomberg, A. (2000) Metabolic surprises in *Saccharomyces cerevisiae* during adaptation to saline conditions: questions, some answers and a model. *FEMS Microbiol Lett*, **182**, 1-8.

Erasmus, D.J., van der Merwe, G.K. and van Vuuren, H.J. (2003) Genome-wide expression analyses: Metabolic adaptation of *Saccharomyces cerevisiae* to high sugar stress. *FEMS Yeast Res*, **3**, 375-399.

Estruch, F. (2000) Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast. *FEMS Microbiol Rev*, **24**, 469-486.

Gancedo, C. and Serrano, R. (1989) Energy-yielding metabolism. In Rose, A.H. and Harrison, J.S. (eds). *The Yeasts*, Vol. 3, NY: Academic Press.

Lin, Y.-H., Bayrock, D.P. and Ingledew, W.M. (2002) Evaluation of *Saccharomyces cerevisiae* grown in a multi-stage chemostat environment under increasing levels of glucose. *Biotechnol Lett*, **24**, 449-453.

Ruis, H. and Schuller, C. (1995) Stress signaling in yeast. *Bioessays*, **17**, 959-965.

Teusink, B., Walsh, M.C., van Dam, K. and Westerhoff, H.V. (1998) The danger of metabolic pathways with turbo design. *Trends Biochem Sci*, **23**, 162-169.

Thomas, K.C., Hynes, S.H., Jones, A.M. and Ingledew, W.M. (1993) Production of fuel alcohol from wheat by VHG technology: effect of sugar concentration and fermentation temperature. *Appl Biochem Biotech*, **43**, 211-226.

Thomas, K.C., Hynes, S.H. and Ingledew, W.M. (1994) Effects of particulate materials and osmoprotectants on very-high-gravity ethanolic fermentation by *Saccharomyces cerevisiae*. *Appl Environ Microbial*, **60,** 1519-1524.

Thomas, K.C. and Ingledew, W.M. (1992) Production of 21% (v/v) ethanol by fermentation of very high gravity (VHG) wheat mashes. *J Ind Microbiol,* **10**, 61-68.

Zhao, Y. and Lin, Y.-H. (2003) Growth of *Saccharomyces cerevisiae* in a chemostat under high glucose conditions. *Biotechnol Lett*, **25**, 1151-1154.

5.2.7 Additional experimental details

Protein profiling analysis

To analyze the protein profiles in yeast cells, three steps were generally involved. They were: 1) the preparation of cell-free extracts by means of a cell disruption device, so that mixed protein samples were obtained; 2) the digestion of protein mixtures using trypsin, which enabled complex peptide mixtures samples to be prepared; and 3) the analysis of peptide mixtures using the combination of HPLC and tandem mass spectrometry (MS/MS), such that the peptide sequences so obtained were identified and corresponding proteins were determined by means of protein identification tools. The experimental protocols of these steps are detailed in the following subsections.

Protocol for cell-free extract preparation

Cell pellets were suspended in 3 ml cold (4°C) cell disruption buffer, consisting of 50 mM potassium phosphate (pH 7.4), 2mM $MgCl_2$, 1 mM dithiothreitol (DTT), 2 mM phenylmethylsulphonyl fluoride (PMSF), and 2 mM EDTA. Then cell suspensions were passed three times through a chilled French Pressure Cell (SLM Instruments Inc.

Urbana, IL) at 20,000 psi. The resulted extracts were then centrifuged at 17,000 ×g for 30 min (4°C), the clear supernatant was used for subsequent protein assay using the Bradford method (Sigma, Oakville, ON), in which bovine serum albumin (BSA) was used as the standard reference protein.

Protocol for protein mixture digestion

The extracted protein solutions obtained from above steps were adjusted to 2 mg protein/ml using 8 M urea, and 500 µl samples were placed into a vial. Then 500 µl of 100 mM ammonium bicarbonate buffer (short form: ABB, pH 8.0) were added. After that, 250 µl aliquots of 50 mM DTT were added to the vial to reduce any disulfide bonds in the protein solution allowing the reaction to proceed at 60 °C for 60 min. Then, 250 µl aliquots of 100 mM iodoacetamide solution were added to carboxyamidomethylate all cysteine residues in the protein solution. This was allowed to react in the dark at room temperature for 30 min. After that 10 µl aliquots of 100 mM $CaCl_2$ and 10 µl aliquots of 2 mg/ml trypsin (sequencing grade, Roche Applied Sci., Laval, QC) were added to the solution, and allowed to react at 37 °C for 24 hours to digest proteins and generate peptide mixtures. Finally the pH of the solution was adjusted to 2.7-3.0 and samples were stored in the freezer for further analysis.

Protein identification using MudPIT

The tryptic peptide mixtures obtained from above steps were then analyzed by an off-line multidimensional LC-MS/MS system for protein identification. This process was

subdivided into four steps: 1) fractionation and collection of peptide mixtures using strong cation ion exchange column (SCX); 2) desalting of fractionated peptide mixtures; 3) separation of the desalted peptide mixtures using reverse phase HPLC and analysis of each separated peptide using a tandem mass spectrometer; and 4) interpretation of the obtained mass spectral data using two protein identification tools (Mascot and our two-pass approach), the identified proteins from each tool were then analyzed by our developed two-step strategy to obtain the proteins identified with high confidence. The detail information of these steps is described in the following subsections.

Fractionation and collection of peptide mixtures

Each tryptic peptide mixture obtained from the digestion protocol was loaded onto a 2.1×100 mm polysulfoethyl A column (POLYLC Inc, Columbia, MD), which was connected to an HPLC (Model 1100 series, Agilent Technologies, CA). Three buffer solutions were used during the 80-min gradient separation at a flow rate of 0.2 ml/min. These buffers include: A) 5 mM KH_2PO_4 (pH 3.0), mixed with 25% (v/v) acetonitrile (ACN); B) 5 mM KH_2PO_4 (pH 3.0) and 0.25 M KCl, mixed with 25% (v/v) ACN; and C) 5 mM KH_2PO_4 (pH 3.0) and 0.5 M KCl, mixed with 25% (v/v) ACN. The linear gradient condition was: 0-10 min, 100% A; 10-64 min, 0 to 100% B along with 100 to 0% A; 64-80 min, 0-100% C along with 100 to 0% B. The fraction collected during the first 10 min was considered as the zero-time sample (Fraction 0). After that, fractions were collected every seven minutes. Therefore, a total of 11 fractions for each peptide mixture after SCX separation were collected.

Desalting of fractionated peptide mixtures

To prevent KCl from plugging the tandem mass analyzer, each fractionated peptide mixture sample was subjected to a desalting process using MiniSpin silica C_{18} column kits (Vydac Inc., Hesperia, CA) according to the manufacturer's instructions. Brifely speaking, this process is composed of three steps: 1) conditioning the column using 100% ACN plus centrifugation; 2) processing sample by centrifugation; and 3) releasing the sample by washing the sample using 80% ACN.

LC-MS/MS analysis

Trypsinized *S. cerevisiae* peptide mixtures (one sample is from the cells grown at 10 and the other sample is from the cells grown at 300 g glucose/l, 11 fractions for each sample) were analyzed by LC-MS/MS at the Plant Biotechnology Institute, Saskatoon, SK, Canada. LC-MS/MS analysis was performed using a capLC pump interfaced to a Q-TOF Ultima global hybrid tandem mass spectrometer fitted with a Z-spray nanoelectrospray ion source (Micromass, Waters, MA). Each fraction of peptide mixture was loaded onto a C_{18} trapping column (Symmetry ^{TM} 300, 0.35×5 mm Opti-pak; Waters, MA) and washed for 3 min using solvent C (Milli-Q grade water with 0.2% formic acid) at a flow rate of 30 μl/min. The flow path was then switched using a 10-port rotary valve, and the sample eluted onto a C_{18} analytical column (PepMap^{TM}, 75 μm×15 cm, 3-μm particle size; LC Packings, Waters, MA). Separations were performed using a linear gradient of 0% to 65% solvent A (A: acetonitrile with 0.2% formic acid) over 70 min. The composition was then changed to 80:20% of A:B (Solvent B: water

with 0.2% formic acid, the same as Solvent C) and held for 10 min to flush the column before re-equilibrating for 7 min at 100% of B. Mass calibration of the Q-TOF instrument was performed using a product ion spectrum of Glu-fibrinopeptide B acquired over the *m/z* range 50 to 1900 Da. LC-MS/MS analysis was carried out using data dependent acquisition, during which peptide precursor ions were detected by scanning from *m/z* 400 to 1200 Da in TOF MS mode. Multiple charged (+2, +3, or +4) ions rising above predetermined threshold intensities were automatically selected for TOF MS/MS analysis, and product ion spectra were acquired over the *m/z* range of 50 to 1950 Da. Each fraction of collected sample results in a PKL file (Figure 5.5) after LC-MS/MS analysis, and the mass spectral data (monoisotopic form) was processed using ProteinLynx software (Micromass, Waters, MA).

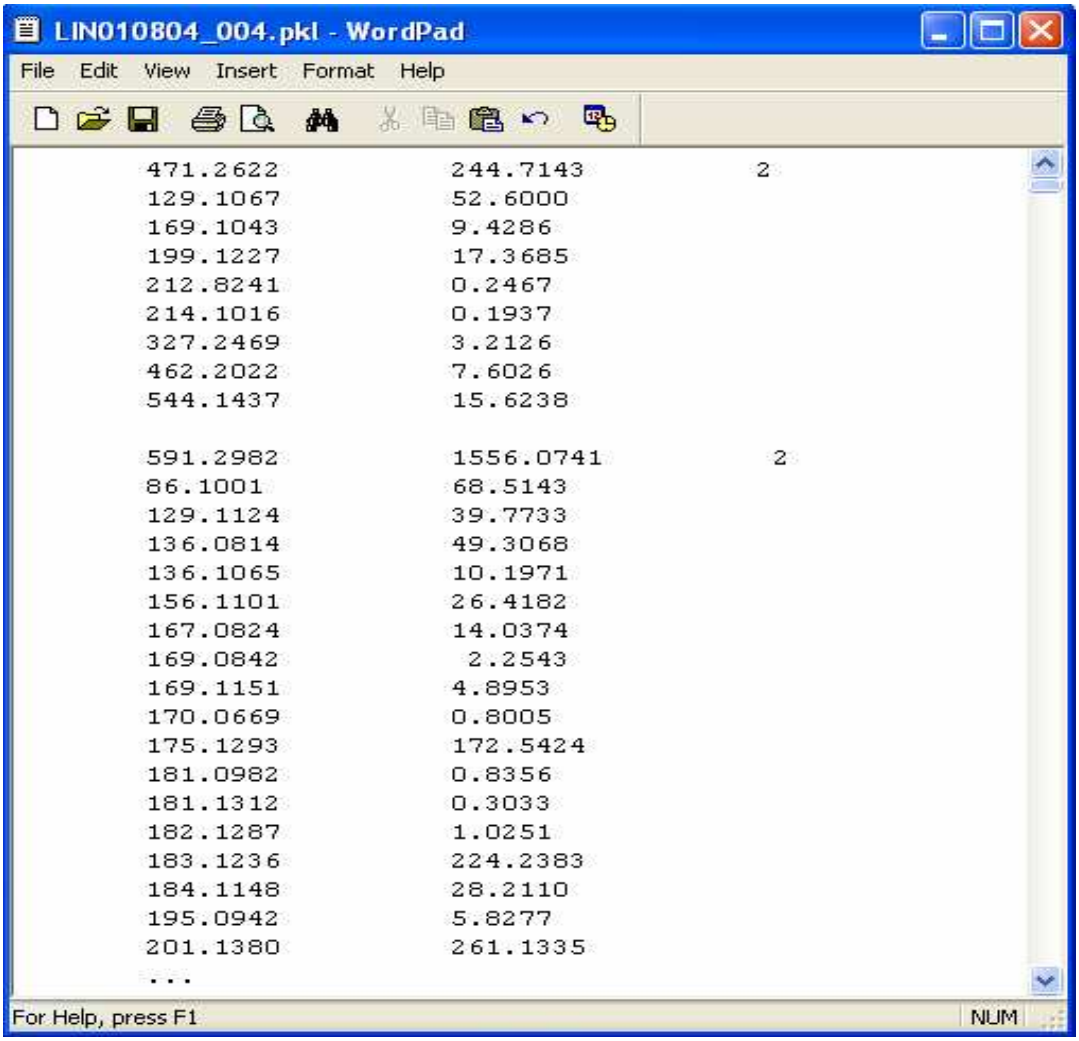


Figure 5.5 Partial listing of a PKL file (Fraction 4, 300 g glucose/l)
In each PKL file, many records are separated by an empty space. Each record gives the information of a precursor ion and its corresponding product ion series. For example, the first row in each record (e.g., the first record) is composed of the mass data of a precursor ion (471.2622 Da), its intensity (244.7143) and charge state (2); the data listed from the second row to the last row (from row 2 to row 9) in the record give the mass data of the dissociated product ion series (first column) and their intensities (second column) corresponding to the precursor ion of 471.2622 Da.

Interpretation of mass spectral data

The experimentally obtained PKL files were fed to both Mascot software (Matrix Science Ltd., London, UK) and our two-pass approach for protein identification. The method for protein identification by Mascot was briefly introduced in Chapter 3 and the method used in our two-pass approach was described in Chapter 4. The criteria used for protein search were: 1) only monoisotopic ions were searched for the precursor ions and product ions; 2) the maximum missed cleavage was 1; 3) the charge states of searched peptides were +2, +3, and +4; 4) the searched mass tolerance for precursor ion and its corresponding product ion series were 1.0 and 0.8 Da, respectively; and 5) the standard *S. cerevisiae* protein database used in searching process was the SwissProt Protein Database (Version 42.6).

The identified proteins from the above two protein identification tools were analyzed to locate the proteins with the highest confidence using our developed two-step strategy as described in Chapter 3. Briefly speaking, the peptides identified from each protein identification tool were grouped into unique peptides and non-unique peptides for protein identification. The proteins identified by unique peptides were considered as the high confidence proteins. Then the high confidence proteins identified from each tool were cross-compared to locate the common proteins, which were then classified as the proteins with the highest confidence. The highest confidence proteins were used in the subsequent data interpretation and discussion.

Chapter 6 Conclusions and future work

6.1 Generation discussion

Currently, mass spectrometry is the major means for proteomic study. The interpretation of the resultant MS spectral data is a crucial task. Since each protein identification tool has its own logic and limitation, the proteins identified using different software packages may vary from one to another, such that the confidences in identification of proteins by various tools may vary significantly. In other words, the protein identification process is still an ‘art’, thus attention should be paid when drawing conclusions based on these ‘identified’ proteins. Since there is no standard MS database for all proteins of interest, most current software tool developers suggest to MS scientists or proteomic researchers to manually assess the final identification report. The curation process, however, is a time-consuming process, especially for those conducting whole-cell proteome analyses.

The two-step strategy developed in our laboratory, integrating the unique peptide characteristics and combining cross comparison analysis, classifies the levels of confidence of the identified proteins into 4 groups. The Level 4 group of identified proteins was regarded with the highest confidence; proteins in this group can be used to draw conclusions with confidence.

The proposed two-pass tool was developed based upon experimental conditions employed in the LC-MS/MS runs. By providing the m/z data of precursor ion selected from an MS spectrum, a list of possible peptides (falling within the specified m/z ± the allowable mass tolerance) was obtained. As a result, the size of the peptide database was reduced dramatically (hereafter called reduced peptide database). With further provision of fragment ions coming from the same precursor ion, the second pass of the approach identified the most probable peptides using the reduced peptide database on the basis of the total number of matched product ions, and the total number of matched b- and y-ions for each peptide candidate. The proposed approach is not only capable of identifying the most probable peptide, but also gives detailed identification information for later confirmation; and this is the obvious advantage of our developed tool. In addition, the proposed approach is very easy to implement by proteomic scientists to carry out protein identification tasks. All that is required is the MS and MS/MS data and the standard protein database of species of interests (for generating a peptide database). In fact, no prior MS knowledge is required to operate this two-pass searching protein identification tool.

When *S. cerevisiae* was grown in a chemostat, it was found that the ethanol production yield at low glucose concentration was higher than that at high glucose concentration even though the high glucose concentration favored ethanol production (total ethanol concentration was higher). The protein profile suggested that more nutrient (sugar) was channelled into the PP pathway when *S. cerevisiae* was grown under a high glucose concentration. The reason for this phenomenon might be that the cell needs more

reducing power (NADPH) for the synthesis of macromolecules such as proteins, nucleic acids, and lipids. These materials are essential for the cell to modify its structure (cell wall) in order to survive osmotic stress and to replicate.

Single-stage continuous fermentation can only utilize a fraction of the carbon source supplied (e.g., in this study, about 54% of glucose was utilized for ethanol synthesis, resulting in a relatively lower $Y_{p/s}$ than in batch operation). In contrast, a multi-stage continuous fermentation converts more sugar to ethanol and maintains a relatively high apparent $Y_{p/s.}$ Therefore, a multi-stage continuous operation seems to be a promising alternative for ethanol fermentation under high specific gravity conditions. Besides the fermentation mode, modification of medium is necessary for ethanol production under very high gravity conditions. For example, proline, an osmotic-protection chemical, can be added in the fresh medium to help cell grow and increase the number of survival cells. Thus a resultant of high ethanol production yield could be achieved.

6.2 Conclusions

I. The developed two-pass protein identification tool can identify proteins and gives detailed information.
II. The developed two-step strategy can classify the identified proteins into different levels of confidence.
III. The glucose concentration in the nutrient feed affected significantly the production of ethanol and the protein profilings in the cell.

IV. By comparing the proteins in the central metabolic pathways, we postulate that under the high gravity condition:
    a. the PP pathway was highly activated to maintain cell life,
    b. more glucose was channelled into the PP pathway to generate NADPH,
    c. the enzymes CIT1 and PYC were expressed to provide precursors for proline synthesis.

6.3 Future work

6.3.1 Development of protein or peptide enrichment techniques

Proteomic study includes protein(s) and/or peptide(s) separation and protein(s) and/or peptide(s) identification. The concentration of proteins obtained after separation greatly affects the subsequent analysis, e.g., the quality of MS spectral data. Although many sample pre-treatment methods have been reported, they have their own advantages and disadvantages when used in conjunction with MudPIT. Thus a suitable protein or peptide enrichment technique must be developed and implemented for future proteomic analyses.

6.3.2 Improvement of our developed tools

In the future, some modifications will be made for our developed protein identification tools. For example, more protein databases of various species will be provided for researchers with different species of interest. Furthermore, a user-friendly graphic operational interface should be developed for our developed tools before posting them to the Internet for public access.

Appendix A The development of an algorithm for the mass spectral interpretation of phosphoproteins

This chapter has been published in *Proteomics*, **5,** 843-845 (2005). Part of the contents in this chapter was presented at the 3rd International Proteomics Conference (IPC'03) at Taipei (Taiwan), May 14-17, 2004.

A.1 Abstract

Extended from the peptide mapping method (Fenyo, 2000), the proposed algorithm takes the mass information of a precursor ion to re-construct all possible phosphorylated peptide sequences. The mass spectra of product ions from the corresponding precursor ion is then used and compared to the re-constructed sequences to deduce the most probable phosphoprotein. The proposed algorithm also predicts all possible combinations of phosphopeptides, which may serve as a clue for designing proper phosphorylation experiments to validate the existence of these peptides and the corresponding proteins.

A.2 Introduction

Serine (S), threonine (T), and tyrosine (Y) are the amino acids that are most often phosphorylated, resulting in the so-called phosphoproteins. Approximately one-third of

mammalian proteins are phosphoproteins. The ratio of phosphorylation for the three different amino acids is approximately 1000/100/1 for S/T/Y (http://www.indstate.edu/thcme/mwking/protein-modifications.html). Phosphorylation is an addition of HPO_3 to the hydroxyl side group of S, T, and Y, resulting in the H_2PO_4 moiety attached to the side-group of a carbon atom. Clearly, there can be more than one phosphate on a protein, and the phosphate moiety can occur on adjacent residue sites or on more widely spaced residues in the protein sequence.

Many experimental approaches have been implemented for identifying phosphorylated proteins, such as Edmen degradation, phosphor-labelling, immunoprecipitation, etc. These approaches are specific and selective for locating the phosphorylation site of a single protein for each experiment. In contrast, multidimensional protein identification technology that couples HPLC to tandem mass spectrometry can systematically identify all expressed proteins in one run (Peng *et al*., 2003; Washburn *et al*., 2001). The identification of expressed phosphoproteins is particularly important to comprehensively understand the protein function relative to the extraneous environment. Hence, the development of searching algorithms to interpret MS and/or MS/MS spectra has attracted great attention. There are many software tools that have been implemented to identify unmodified proteins based on the fragmented MS spectra; including Mascot (Perkins *et al*., 1999), PepFrag (Qin *et al*., 1997), MS-Tag (Clauser *et al*., 1999), PepSea (Mann and Wilm, 1994) and SEQUEST (Eng *et al*., 1994). For the identification of phosphoproteins, however, the related tools are still limited.

Extended from the peptide mapping method, we propose an algorithm that takes MS data and subsequently generates all possible combinations of phosphorylated peptide sequences. By incorporating MS/MS spectra with newly generated peptide sequences, the most probable phosphopeptide can be searched, resulting in the identification of the corresponding phosphoprotein. Data extracted from literature (Synder, 2000; Wu *et al*., 2003) was used to validate the proposed algorithm.

A.3 Methods

In addition to the peptide mapping method, the number of phosphates attached to a protein sequence and the site of phosphorylation on the sequence are two key issues which must be addressed when developing a phosphorylation-searching algorithm. In this proposed algorithm, an *in silico* tryptic digested peptide database was generated using the protein database retrieved from publicly accessible resources such as the Kyoto Encyclopaedia of Genes and Genomes (http://www.genome.jp/kegg/). By providing the experimentally obtained m/z data and the charge state (z) of a precursor ion, and searching through the *in silico* peptide database, a list of matched candidate peptide(s) with predicted mass (MW) was obtained, the number (n) of S, T, and Y was counted, and the number (m) of phosphate moiety (HPO_3) attached to the candidate peptide was estimated from the formula: $m/z = \frac{MW + z + 80m}{z}$. All the possible combinations of phosphorylated peptides were then deduced according to $C_m^n = \frac{n!}{(n-m)!m!}$, and were subsequently used and compared to the product ions series data from the experiment. The most probable phosphorylated peptide was obtained, and

the corresponding phosphoprotein was thus determined. The criterion for the selection of the most probable phosphopeptide is based on the peak ratio, which is defined as the ratio of the number of matched *in silico* product ions to the number of product ions from the experiment. A peak ratio of 1 indicates that a complete match is obtained.

A.4 Implementation

To validate the proposed algorithm, peptide information retrieved from Synder (2000) and Wu *et al*. (2003) was collected to form a pooled peptide database (Table A.1). The fact that this peptide database was used instead of using procedures described above was due to the following: 1) the experimental data used for validation was extracted from different species; and 2) the sequences in the pooled database were not all terminated with lysine (K) or arginine (R) because different proteolytic enzyme systems were used. In this database, the actual name of a respective protein used in References 9 and 10 was substituted by ‘protein1’, ‘protein2’, to ‘protein5’; and only one peptide sequence corresponding to each of these assigned proteins was listed.

Table A.1 Pooled peptide database*

Protein	MW of peptide		Sequence of peptide
protein1:	[1]	1416.64157	[1]ISHEIESSSSEVN[13]
///			
protein2:	[1]	2421.34717	[1]LCDFGVSGQLIDSMANSFVGTR[22]
///			
protein3:	[1]	1514.62486	[1]KDSDDEEEVVHVD[13]
///			
protein4:	[1]	2011.04708	[1]DNRSQVETEDLILKPGVV[18]
///			
protein5:	[1]	2120.97873	[1]EKKEFLEPDSWETLDQQ[17]
///			

* The first column is a pseudo protein name (e.g., protein1); the second column is molecular weight of one of the protein's peptides; the third column is the amino acid sequence of one of the protein's peptides, the numbers between the amino acids represent the starting and ending position of this peptide in the original protein sequence.

The proposed algorithm takes mass data of precursor ions from Table A.2, and searches through the pooled database, in which a unit mass tolerance (i.e., 1 Da) was allowed. As an illustration, given a precursor ion with m/z = 861.6 and z = 3, only one possible peptide was found from the pooled database. In this connection, four possible phosphorylated amino acids were identified, and two phosphate groups were estimated (Figure A.1a), resulting in six possible phosphorylated peptides (Figure A.1b). The searched peptide is 22 amino acids in length, positioning from 1 to 22 within ‘protein2’. The lowercase ‘p’ represents the site of phosphorylation. A complete listing of searched results of four tested precursor ions and all possible combinations of phosphopeptides of each respective precursor ion is presented in Sections A and B of Appendix D.

Table A.2 Experimental MS spectral data* retrieved from Synder (2000) and Wu *et al*. (2003)

Precursor ions (m/z, z)	Product ions series (m/z)
798.31, +2	370.15, 411.15, 469.22, 526.17, 568.29, 641.20, 682.76, 697.33, 732.29, 770.24, 826.37, 899.29, 955.41, 1028.33, 1070.44, 1127.40, 1185.47, 1226.47, 1352.50, 1363.52
861.6, +3	432.48, 507.08, 549.66, 563.66, 579.66, 620.23, 648.72, 740.89, 746.81, 762.82, 833.90, 860.81, 931.89, 1013.15, 1063.08, 1126.31, 1138.15, 1149.24, 1230.14, 1239.46, 1247.24, 1262.40, 1345.23, 1360.40
1101.48, +2	390.15, 604.28, 650.56, 904.48, 1116.55, 1201.46, 1298.51, 1427.56, 1469.67, 1598.71, 1687.71, 1812.84, 1927.86
1046.52, +2	371.21, 386.17, 499.30, 612.39, 681.26, 725.47, 780.32, 838.55, 909.37, 953.58, 1082.62, 1139.46, 1183.67, 1254.48, 1312.72, 1411.78, 1480.65, 1539.84, 1593.74, 1721.83

* Precursor ion is shown as mass-to-charge ratio (m/z) followed by charge state. All m/z values of both precursor ion and product ion series are monoisotopic mass.

(a)

The possible peptides at m/z = 861.6 are:
protein2_[1]LCDFGVSGQLIDSMANSFVGTR[22] (z=3) (num_of_phos=2)

(b)

1	protein2_[1]LCDFGVSpGQLIDSpMANSFVGTR[22]	(z=3)
2	protein2_[1]LCDFGVSpGQLIDSMANSpFVGTR[22]	(z=3)
3	protein2_[1]LCDFGVSpGQLIDSMANSFVGTpR[22]	(z=3)
4	protein2_[1]LCDFGVSGQLIDSpMANSpFVGTR[22]	(z=3)
5	protein2_[1]LCDFGVSGQLIDSpMANSFVGTpR[22]	(z=3)
6	protein2_[1]LCDFGVSGQLIDSMANSpFVGTpR[22]	(z=3)

(c)*

1 The possible peptide is: protein2_[1]LCDFGVSGQLIDSpMANSpFVGTR[22] (z=3)
The peak ratio is 0.708.

The identified sequences are:

740.89 -->	740.55677 --> b[1--6] [z=1]	LCDFGV
1013.15 -->	1012.66884 --> b[1--9] [z=1]	LCDFGVSGQ
1126.31 -->	1125.75290 --> b[1--10] [z=1]	LCDFGVSGQL
1126.31 -->	1125.59640 --> b[1--19] [z=2]	LCDFGVSGQLIDSpMANSpFV
1239.46 -->	1238.83696 --> b[1--11] [z=1]	LCDFGVSGQLI
507.08 -->	506.83839 --> b[1--9] [z=2]	LCDFGVSGQ
507.08 -->	507.63712 --> b[1--13] [z=3]	LCDFGVSGQLIDSp
563.66 -->	563.38042 --> b[1--10] [z=2]	LCDFGVSGQL
620.23 -->	619.92245 --> b[1--11] [z=2]	LCDFGVSGQLI
1345.23 -->	1344.54137 --> y[1--11] [z=1]	RTGVFSpNAMSpD
1230.14 -->	1229.51488 --> y[1--10] [z=1]	RTGVFSpNAMSp
1149.24 -->	1149.51488 --> y[1--10] [z=1]	RTGVFSpNAMS
1063.08 -->	1062.48285 --> y[1--9] [z=1]	RTGVFSpNAM
931.89 -->	931.44236 --> y[1--8] [z=1]	RTGVFSpNA
860.81 -->	860.40525 --> y[1--7] [z=1]	RTGVFSpN
860.81-->	861.46209 --> y[1--22] [z=3]	RTGVFSpNAMSpDILQGSVGFDCL
746.81 -->	746.36232 --> y[1--6] [z=1]	RTGVFSp
579.66 -->	579.33029 --> y[1--5] [z=1]	RTGVF
432.48 -->	432.26188 --> y[1--4] [z=1]	RTGV
648.72 -->	647.96862 --> y[1--17] [z=3]	RTGVFSpNAMSpDILQGSV

///

Figure A.1 (a) Illustrated search results for the precursor ion at m/z = 861.6 and z = 3, (b) Listing of all possible combinations of the phosphopeptide at m/z = 861.1 and z =3, (c) Listing of the identified most probable phosphopeptide at m/z = 861.6 and z =3.
*Column 1, experimentally obtained mass of product ions series; Column 2, predicted mass of product ions series; Column 3, ion type; Column 4, predicted charge state; Column 5, predicted peptide sequence.

These six possible phosphorylated peptides were then fragmented *in silico*, resulting in the generation of a series of b-type and y-type product ions for each respective peptide. The mass data of these *in silico* product ions for each possible phosphopeptide was matched to the experimentally obtained mass data of the product ions (Table A.2). A phosphopeptide that has the highest peak ratio value was regarded as the most probable one, such that the corresponding phosphoprotein was considered identified. As seen in Figure A.1c, the identified phosphopeptide is 'LCDFGVSGQLIDSpMANSpFVGTR' as previously reported by Synder (2000), and the corresponding phosphoprotein is 'protein2'. The site of phosphorylation is at 13 and 17 of the identified phosphopeptide. A complete listing of searched results of four tested peptides can be found in Section C of Appendix D.

A.5 Concluding remarks

The proposed algorithm is experiment-oriented. Given the genome sequence of the species of interest, mass information of precursor ions, product ions series, and the charge state, the most probable phosphopeptide, and the corresponding phosphoprotein is identified. Owing to the specific sites of phosphate groups attached to a peptide, the use of peak ratio values as the identification criterion is confident; the more phosphate groups attached, the higher the level of confidence of the identified phosphoprotein. The proposed algorithm is easy to implement, and can be readily extended to identify proteins subjected to other PTM such as acetylation, methylation, and so on.

A.6 References

Clauser, K.R., Baker, P. and Burlingame, A.L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem*, **71**, 2871-2882.

Eng, J.K., McCormack, A.L. and Yates, J.R., III. (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.

Fenyo, D. (2000) Identifying the proteome: software tools. *Curr Opin Biotechnol*, **11**, 391-395.

Mann, M. and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem*, **66**, 4390-4399.

Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. and 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.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551-3567.

Qin, J., Fenyo, D., Zhao, Y., Hall, W.W., Chao, D.M., Wilson, C.J., Young, R.A. and Chait, B.T. (1997) A strategy for rapid, high-confidence protein identification. *Anal Chem*, **69**, 3995-4001.

Synder, P.A. (2000) *Interpreting Protein Mass Spectra: A Comprehensive Resource*. New York, NY: Oxford University Press.

Washburn, M.P., Wolters, D. and Yates, J.R., III. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, **19**, 242-247.

Wu, C.C., MacCoss, M.J., Howell, K.E. and Yates, J.R., III. (2003) A method for the comprehensive proteomic analysis of membrane proteins. *Nat Biotechnol*, **21**, 532-538.

Appendix B An automated approach to extract metabolically related proteins for metabolic flux analysis of *Pseudomonas putida*

This chapter was presented and published as part of *Proceeding* (WE-191, 1-7) at the 1st Water and Environment Specialty Conference of the Canadian Society for Civil Engineering at Saskatoon (Canada), June 2-5, 2004.

B.1 Abstract

Pseudomonas putida has been widely used to treat environmental pollutants. This species, like some other microorganisms, utilizes and degrades hazardous substances for growth. To design effective bioremediation processes, a comprehensive understating of how *P. putida* responds to extraneous disturbances are essential. One approach to obtain a global viewpoint of cellular work is through metabolic flux analysis. The analysis requires the construction of a metabolic pathway network, which is interwoven by metabolites and the related enzymes (proteins). Thus, the protein expression profile of a species grown under specific conditions becomes indispensable for constructing a physiologically meaningful metabolic pathway network. In this paper, an automated approach was proposed to utilize the publicly accessible genome information for the above-stated purposes. The proposed approach retrieves intended bio-information stored at three different databases, and combines them to construct a metabolic pathway

network. The network can then be used to explore and identify possible intracellular reaction bottlenecks during bioremediation, which could then be utilized for subsequent strain improvement to enhance cellular survivability under a harsh environment, biodegradation capability and efficacy.

B.2 Introduction

Pseudomonas putida grows on a wide variety of different types of organic compounds using a broad array of metabolic pathways (Wackett, 2003), and has been considered as a versatile biocatalyst for processing environmental pollutants such as the aromatic chemicals benzene, toluene, xylene, and related compounds. In nature, this organism is commonly propagated under harsh conditions, such that *P. putida* suffers various stresses from the surrounding environment (Estruch, 2000; Ruis and Schuller, 1995); consequently impacting its efficiency in wastewater treatment. To gain information relating to the intracellular work for designing effective bioremediation processes, a global view of physiological variation of *P. putida* relating to extraneous environment is needed.

With the aid of a metabolic pathway network, metabolic flux analysis (MFA), a method of exploring cellular physiology in a global perspective, is used to estimate the distribution of nutrient flux throughout the whole cell system. In MFA, by considering the high turnover of metabolite pools, the intracellular metabolites are assumed at a pseudo-steady state, and the associated bio-reactions are used to construct a metabolic pathway network, resulting in a set of linear metabolite balancing equations. Generally,

a linear programming technique is used to obtain a feasible solution for such an under-determined problem. The solution so obtained includes the metabolic flux of each reaction in the network and the estimated metabolite concentrations. The magnitude of the metabolic flux of one reaction indirectly reflects the strength of the enzyme activity, such that possible reaction bottlenecks can be identified by examining the magnitude of a metabolic flux (Zhao and Lin, 2002). Most importantly, from the network point of view, the cellular functionality under a known growth environment can be elucidated. Since enzymes and structural proteins are products of genes, by knowing the magnitude of metabolic flux, one could reason which gene or a group of related genes is being under- or over-regulated. As a result, genetic engineering techniques can be applied to improve strain performance. For instance, metabolic pathway engineering (MPE) focuses on the manipulation of hundreds of different genes simultaneously, preventing cells from secreting environmentally harmful compounds, and generating potentially important and commercially valuable products for customer-driven needs during the waste treatment process.

Two factors are crucial for an accurate estimation of metabolic fluxes throughout the cell. They are reactions and metabolite concentrations, among which the choice of reactions is more important. This information is generally gathered from literature and acceptable hypotheses (Stephanopoulos *et al*., 1998) but are not detailed or complete. Recently, the genome sequence analyses of a number of organisms (e.g., *P. putida*, (Nelson *et al*., 2002; Stjepandic *et al*., 2002; Weinel *et al*., 2002) have been completed and are publicly available. Additionally, a number of bioinformatics databases are also accessible through the website, making it possible to reconstruct a complete metabolic

network. To facilitate and automate the construction of a metabolic pathway network, we propose an approach to inter-relate these databases and extract required information for MFA and MPE analyses.

B.3 Strategy and implementation

The proposed automated approach can be divided into four steps in conjunction with three publicly available databases (Figure B.1). The detailed processing procedures are described below. Partial listing of results after each step is presented for demonstration. The complete outputs can be obtained upon request.

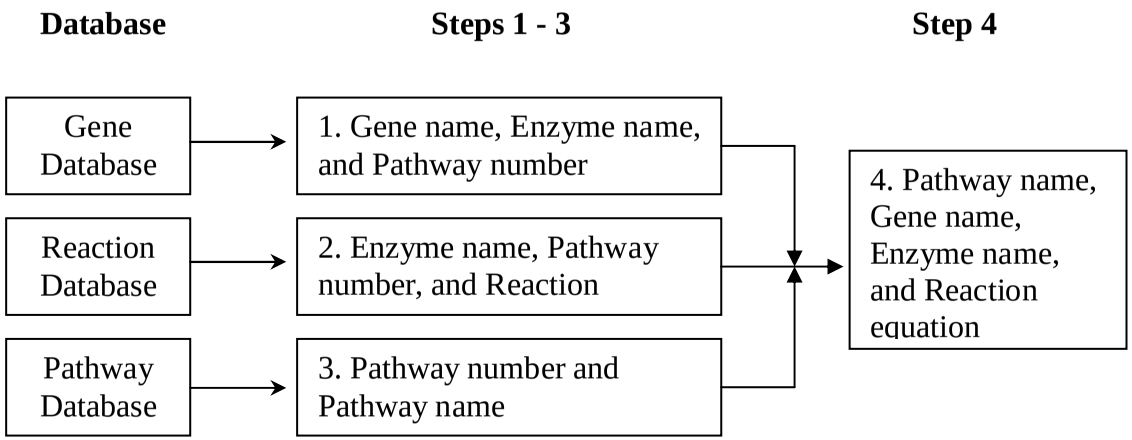


Figure B.1 Schematic of proposed automated approach

B.3.1 Gene database and Step 1

The complete gene database (P.putida.ent,) of *Pseudomonas putida* was downloaded from KEGG (Kyoto Encyclopaedia of Genes and Genomes) at the following website ftp://ftp.genome.ad.jp/pub/kegg/genomes/genes/P.putida.ent (March 03, 2004). Each gene defined in the database contains many attributes, and the overall attributes of a gene is called a record. A sample record contained in the database is shown in Figure B.2. Since MFA requires the information of reactions, while MPE needs the genes and the related proteins (enzymes), only the gene ID, gene name, enzyme name, and pathway number were extracted using PERL language. A partial listing of output is attached in Figure B.3. The first column 'pp****' of the figure refers to gene ID, the second column is the name of the gene, the output beginning with [EC:*.*.*.*] is the enzyme name, followed by the metabolic pathway which this enzyme is involved in.

```
ENTRY       PP0011            CDS       P.putida
NAME        dnaN
DEFINITION  DNA polymerase III, beta subunit [EC:2.7.7.7]
ORTHOLOG    KO: K02338  DNA polymerase III beta subunit
CLASS       Metabolism; Nucleotide Metabolism; Purine metabolism
[PATH:ppu00230]
            Metabolism; Nucleotide Metabolism; Pyrimidine metabolism
            [PATH:ppu00240]
            Genetic Information Processing; Replication and Repair; DNA
            polymerase [PATH:ppu03030]
POSITION    11103..12206
DBLINKS     TIGR: PP0011
            NCBI: 24981347
            SPTR: P13455
CODON_USAGE       T                 C                 A                 G
          T   1   8   0  10   1   7   1   5   1   6   1   0   0   3   0   0
          C   3   1   0  38   5   3   0   8   2   3   7  14   5  19   2   0
          A   4  13   0   9   6  12   1   1   2  10   6   8   1  11   0   0
          G   4  13   6   9   3  12   1   7   7  12  15  17  12  11   1   0
AASEQ       367
            MHFTIQREALLKPLQLVAGVVERRQTLPVLSNVLLVVQGQQLSLTGTDLEVELVGRVQLE
            EPAEPGEITVPARKLMDICKSLPNDALIDIKVDEQKLLVKAGRSRFTLSTLPANDFPTVE
            EGPGSLTCNLEQSKLRRLIERTSFAMAQQDVRYYLNGMLLEVSRNTLRAVSTDGHRLALC
            SMSAPIEQEDRHQVIVPRKGILELARLLTDPEGMVSIVLGQHHIRATTGEFTFTSKLVDG
            KFPDYERVLPKGGDKLVVGDRQALREAFSRTAILSNEKYRGIRLQLAAGQLKIQANNPEQ
            EEAEEEISVDYEGSSLEIGFNVSYLLDVLGVMTTEQVRLILSDSNSSALLQEAGNDDSSY
            VVMPMRL
NTSEQ        1104
            atgcatttcaccattcaacgcgaagccctgttgaaacccctgcaactggtcgccggtgtc
            gtcgagcgccgtcagaccttgccggtcctgtccaacgtattgctggtcgtacaaggccag
            caactgtcgttgaccggtaccgacctggaagtcgagctggtaggccgcgtgcaactggaa
            gagccggccgagcctggcgagatcactgttcctgcgcgcaagctgatggacatctgcaaa
            agcctgcctaacgacgccctgatcgatatcaaggtcgatgagcagaaattgttggtaaag
            gccggccgcagccgcttcaccctgtcgaccctgcctgccaatgacttcccgactgtggaa
            gaaggcccgggttcgctgacctgcaacctggagcagagcaagctgcgccgcctgatcgag
            cgaaccagcttcgccatggcccagcaggatgtgcgctattacctcaacggcatgctgctt
            gaggtttcccgcaacacgttgcgcgctgtatccaccgacggtcaccgcctggcgctttgc
            tcgatgagcgcgccgatcgagcaggaggatcgacaccaggtcatcgtgccgcgtaaaggt
            atcctggagttggcgcgcctgctgaccgatcctgaaggtatggtcagcatcgtgctgggc
            cagcatcacatccgcgccactactggtgagttcacctttacctccaagctggtggacggc
            aaattcccggactacgagcgcgtactgcccaaaggcggtgacaagctggtcgtcggagac
            cgccaggcgctgcgtgaagccttcagccgtacagcgattctttccaacgaaaagtaccgc
            ggtattcgcctgcagctggcagccggtcaactgaagatccaggctaacaaccccgagcag
            gaagaagccgaagaagaaatcagcgtggactacgaaggtagctcgctggagattggtttc
            aacgtcagctacttgctggacgtgctgggcgtgatgactactgagcaagttcgtctgatc
            ttgtccgattcaaacagcagtgcgctgctgcaggaagctggcaatgacgattcttcctac
            gttgtcatgccgatgcgcctgtaa
//
```

Figure B.2 A partial listing of a gene database for *P. putida*

```
PP0001 parB
PP0002
PP0003 gidB [EC:2.1.-.-]
PP0004 gidA
PP0005 trmE
PP0006 yidC [PATH:ppu03060]
PP0007
PP0008 rnpA [EC:3.1.26.5]
PP0009 rpmH [PATH:ppu03010]
PP0010 dnaA
PP0011 dnaN [EC:2.7.7.7] [PATH:ppu00230] [PATH:ppu00240] [PATH:ppu03030]
PP0012 recF
PP0013 gyrB [EC:5.99.1.3]
......
```

Figure B.3 A partial listing after Step 1. The first column ‘pp****’ refers to gene ID; the second column is the name of the gene; the output beginning with [EC:*.*.*.*] represents the enzyme EC number corresponding to the gene, [PATH: ***] indicates the pathway that the gene is involved in.

B.3.2 Reaction database and Step 2

The reaction database (reaction) of *P. putida* was downloaded from KEGG at the website ftp://ftp.genome.ad.jp/pub/kegg/ligand/reaction (March 03, 2004). Like the gene database, each record contains ENTRY (reaction number), NAME, and DEFINITION as shown in Figure B.4. Only the metabolic pathway number, reaction number, enzyme involved, and the reaction equation were extracted. A partial listing of the output after Step 2 can be found in Figure B.5. The first column of the figure refers to the metabolic pathway number; the second column is the reaction number, followed by the enzyme and stoichiometric equation columns. In the stoichiometric equation column, 'C*****' refers to compounds involved in the equation, e.g., C04184 represents 4-hydroxy-4-methyl-2-oxoglutarate. The complete listing of compound defined in KEGG can be found at its website (ftp://ftp.genome.ad.jp/pub/kegg/ligand/compound).

```
ENTRY       R00006
NAME        2-Acetolactate pyruvate-lyase (carboxylating)
DEFINITION  2-Acetolactate + CO2 <=> 2 Pyruvate
EQUATION    C00900 + C00011 <=> 2 C00022
PATHWAY     PATH: RN00770  Pantothenate and CoA biosynthesis
ENZYME      2.2.1.6
///
ENTRY       R00007
NAME        4-Hydroxy-4-methyl-2-oxoglutarate pyruvate-lyase
DEFINITION  4-Hydroxy-4-methyl-2-oxoglutarate <=> 2 Pyruvate
EQUATION    C04184 <=> 2 C00022
PATHWAY     PATH: RN00362  Benzoate degradation via hydroxylation
ENZYME      4.1.3.17
///
ENTRY       R00008
NAME        Parapyruvate pyruvate-lyase
DEFINITION  Parapyruvate <=> 2 Pyruvate
EQUATION    C06033 <=> 2 C00022
PATHWAY     PATH: RN00660  C5-Branched dibasic acid metabolism
ENZYME      4.1.3.17
///
......
```

Figure B.4 A partial listing of a reaction database defined in KEGG

00190	R00004	3.6.1.1	C00013 + C00001 <=> 2 C00009
00220	R00005	3.5.1.54	C01010 + C00001 <=> 2 C00011 + 2 C00014
00770	R00006	2.2.1.6	C00900 + C00011 <=> 2 C00022
00362	R00007	4.1.3.17	C04184 <=> 2 C00022
00660	R00008	4.1.3.17	C06033 <=> 2 C00022
……			

Figure B.5 A partial listing after Step 2. From left to right column, it represents pathway index, reaction index, enzyme EC number, and reaction equation, respectively.

B.3.3 Pathway database and Step 3

To obtain the pathway database (ppu.html), a tar file was retrieved from the site ftp://ftp.genome.ad.jp/pub/ kegg/tarfiles/pathway.weekly.last.tar.Z (March 03, 2004) and then de-compressed. Using the developed PERL program, the pathway number and the corresponding pathway name were extracted (Figure B.6).

```
ppu00010  Glycolysis / Gluconeogenesis - Pseudomonas putida
ppu00020  Citrate cycle (TCA cycle) - Pseudomonas putida
ppu00030  Pentose phosphate pathway - Pseudomonas putida
ppu00031  Inositol metabolism - Pseudomonas putida
ppu00040  Pentose and glucuronate interconversions - Pseudomonas putida
ppu00051  Fructose and mannose metabolism - Pseudomonas putida
ppu00052  Galactose metabolism - Pseudomonas putida
ppu00053  Ascorbate and aldarate metabolism - Pseudomonas putida
ppu00061  Fatty acid biosynthesis (path 1) - Pseudomonas putida
ppu00062  Fatty acid biosynthesis (path 2) - Pseudomonas putida
ppu00071  Fatty acid metabolism - Pseudomonas putida
ppu00072  Synthesis and degradation of ketone bodies - Pseudomonas putida
ppu00100  Sterol biosynthesis - Pseudomonas putida
ppu00120  Bile acid biosynthesis - Pseudomonas putida
ppu00130  Ubiquinone biosynthesis - Pseudomonas putida
ppu00150  Androgen and estrogen metabolism - Pseudomonas putida
ppu00190  Oxidative phosphorylation - Pseudomonas putida
……
```

Figure B.6 A partial listing after Step 3. The first column represents the pathway index and the second column refers to the name of the pathway.

B.3.4 Integration output from above steps (Step 4)

After Steps 1-3, the three output files were merged and used to relate metabolically related proteins to genes, and reactions. A partial listing after the completion of Step 4 is shown in Figure B.7. For each record seen in the figure, it contains gene ID, enzyme number, reaction, and stoichiometric equation. Note that the stoichiometric equation in the figure has been reformatted, making it easy to extract stoichiometric coefficients to conduct the subsequent metabolic flux analysis.

```
Pathway: ppu00010 (Glycolysis / Gluconeogenesis - Pseudomonas putida)
1     PP0338     2.3.1.12   R00210  0 = -1_C00022 -1_C00010 -1_C00006 +1_C00024 +1_C00011 +1_C00005
1     PP0338     2.3.1.12   R02569  0 = -1_C00024 -1_C00579 +1_C00010 +1_C01136
2     PP0339     1.2.4.1    R00014  0 = -1_C05125 -1_C00011 +1_C00068 +1_C00022
2     PP0339     1.2.4.1    R00210  0 = -1_C00022 -1_C00010 -1_C00006 +1_C00024 +1_C00011 +1_C00005
2     PP0339     1.2.4.1    R03270  0 = -1_C05125 -1_C00248 +1_C01136 +1_C00068
3     PP0545     1.2.1.3    R00710  0 = -1_C00084 -1_C00003 -1_C00001 +1_C00033 +1_C00004 +1_C00080
3     PP0545     1.2.1.3    R00711  0 = -1_C00084 -1_C00006 -1_C00001 +1_C00033 +1_C00005 +1_C00080
4     PP0553     2.3.1.12   R00210  0 = -1_C00022 -1_C00010 -1_C00006 +1_C00024 +1_C00011 +1_C00005
4     PP0553     2.3.1.12   R02569  0 = -1_C00024 -1_C00579 +1_C00010 +1_C01136
5     PP0554     1.2.4.1    R00014  0 = -1_C05125 -1_C00011 +1_C00068 +1_C00022
5     PP0554     1.2.4.1    R00210  0 = -1_C00022 -1_C00010 -1_C00006 +1_C00024 +1_C00011 +1_C00005
5     PP0554     1.2.4.1    R03270  0 = -1_C05125 -1_C00248 +1_C01136 +1_C00068
6     PP0555     1.2.4.1    R00014  0 = -1_C05125 -1_C00011 +1_C00068 +1_C00022
6     PP0555     1.2.4.1    R00210  0 = -1_C00022 -1_C00010 -1_C00006 +1_C00024 +1_C00011 +1_C00005
6     PP0555     1.2.4.1    R03270  0 = -1_C05125 -1_C00248 +1_C01136 +1_C00068
7     PP1009     1.2.1.12   R01061  0 = -1_C00118 -1_C00009 -1_C00003 +1_C00236 +1_C00004 +1_C00080
8     PP1011     2.7.1.2    R01600  0 = -1_C00002 -1_C00221 +1_C00008 +1_C01172
8     PP1011     2.7.1.2    R01786  0 = -1_C00002 -1_C00267 +1_C00008 +1_C00668
9     PP1362     2.7.1.40   R00200  0 = -1_C00002 -1_C00022 +1_C00008 +1_C00074
10    PP1612     4.2.1.11   R00658  0 = -1_C00631 +1_C00074 +1_C00001
11    PP1616     1.1.1.1    R00754  0 = -1_C00469 -1_C00003 +1_C00084 +1_C00004 +1_C00080
12    PP1777     5.4.2.2    R00959  0 = -1_C00103 +1_C00668
…… 
```

Figure B.7 A partial listing after Step 4. From second to the last column, it represents gene index, enzyme EC number, reaction index, and reaction equation, respectively.

B.4 Discussion and remarks

The proposed approach to extract metabolically related proteins has been described as above. The complete output found in Figure B.7 serves as a template for reconstructing the metabolic pathway network. In reality, organisms, grown at specific conditions, synthesize only essential enzymes (proteins) to maintain their growth and propagation. This indicates that only parts of the enzymes and proteins in the cell's proteome are being expressed. Hence, to construct an accurate metabolic pathway network to portray global cellular physiology, a measurement of protein expression profile is indispensable. At present, two techniques are widely used in the proteomic research: two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) and high performance liquid chromatography (HPLC) coupled with a tandem mass spectrometer (Link *et al*., 1999; Peng *et al*., 2003; Washburn *et al*., 2001).

After obtaining the protein expression profiles, a complete listing of the metabolic pathway network reconstructed from genome and proteome databases is modified accordingly, resulting in a reduced set of the network. In connection with the proper experimental measurements, metabolic fluxes can be estimated and used to describe intracellular work.

Additionally, the proposed approach also summaries the inter-relationships among genes, enzymes and the associated metabolic pathways. As seen in Figure B.8, one can observe that: 1) a gene may modulate the synthesis of more than one enzymes, e.g., gene

pp1777 (or xanA) encodes the synthesis of enzymes EC:5.4.2.2 and EC:5.4.2.8; 2) an enzyme may modulate by more than one gene, e.g., EC:1.2.4.1 is co-ordinately expressed by genes pp0554 and pp0555; and 3) an enzyme may be involved in more than one metabolic pathways, e.g., EC:1.2.4.1 may alter reactions relating to pathways 00010, 00290, 00620, and 00650. One potential application of this inter-relationship is that it can be used to guide the biologists to conduct strain improvement toward specific objectives, such as the resistance to toxic substances and survival under harsh environments.

```
Pathway: ppu00010 (Glycolysis / Gluconeogenesis - Pseudomonas putida)

1   PP0338 aceF [EC:2.3.1.12] [PATH:ppu00010] [PATH:ppu00620]
2   PP0339 aceE [EC:1.2.4.1] [PATH:ppu00010] [PATH:ppu00290] [PATH:ppu00620] [PATH:ppu00650]
3   PP0545 [EC:1.2.1.3] [PATH:ppu00010] [PATH:ppu00053] [PATH:ppu00071] [PATH:ppu00120]
                        [PATH:ppu00280] [PATH:ppu00310] [PATH:ppu00330] [PATH:ppu00340]
                        [PATH:ppu00380] [PATH:ppu00410] [PATH:ppu00561] [PATH:ppu00620]
                        [PATH:ppu00640] [PATH:ppu00650]
4   PP0553 acoC [EC:2.3.1.12] [PATH:ppu00010] [PATH:ppu00620]
5   PP0554 acoB [EC:1.2.4.1] [PATH:ppu00010] [PATH:ppu00290] [PATH:ppu00620] [PATH:ppu00650]
6   PP0555 acoA [EC:1.2.4.1] [PATH:ppu00010] [PATH:ppu00290] [PATH:ppu00620] [PATH:ppu00650]
7   PP1009 gap-1 [EC:1.2.1.12] [PATH:ppu00010] [PATH:ppu00472]
8   PP1011 glk [EC:2.7.1.2] [PATH:ppu00010] [PATH:ppu00052] [PATH:ppu00500] [PATH:ppu00522]
9   PP1362 pykA [EC:2.7.1.40] [PATH:ppu00010] [PATH:ppu00230] [PATH:ppu00620] [PATH:ppu00710]
10   PP1612 eno [EC:4.2.1.11] [PATH:ppu00010] [PATH:ppu00400]
11   PP1616 [EC:1.1.1.1] [EC:1.2.1.1] [PATH:ppu00010] [PATH:ppu00071] [PATH:ppu00120]
                              [PATH:ppu00350] [PATH:ppu00561] [PATH:ppu00620] [PATH:ppu00680]
12   PP1777 xanA [EC:5.4.2.2] [EC:5.4.2.8] [PATH:ppu00010] [PATH:ppu00030] [PATH:ppu00051]
                              [PATH:ppu00052] [PATH:ppu00500] [PATH:ppu00521] [PATH:ppu00522]
                              [PATH:ppu00530]
13   PP1808 pgi-1 [EC:5.3.1.9] [PATH:ppu00010] [PATH:ppu00030] [PATH:ppu00500]
14   PP2426 [EC:1.1.1.2] [PATH:ppu00010] [PATH:ppu00561] [PATH:ppu00930]
15   PP2589 [EC:1.2.1.3] [PATH:ppu00010] [PATH:ppu00053] [PATH:ppu00071] [PATH:ppu00120]
                              [PATH:ppu00280] [PATH:ppu00310] [PATH:ppu00330] [PATH:ppu00340]
                              [PATH:ppu00380] [PATH:ppu00410] [PATH:ppu00561] [PATH:ppu00620]
                              [PATH:ppu00640] [PATH:ppu00650]

......
```

Figure B.8 A partial listing of the inter-relationship among genes, enzymes, and pathways

B.5 References

Estruch, F. (2000) Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast. *FEMS Microbiol Rev*, **24**, 469-486.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M. and Yates, J.R., III. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol*, **17**, 676-682.

Nelson, K.E., Weinel, C., Paulsen, I.T., Dodson, R.J., Hilbert, H., Martins dos Santos, V.A., Fouts, D.E., Gill, S.R., Pop, M., Holmes, M., Brinkac, L., Beanan, M., DeBoy, R.T., Daugherty, S., Kolonay, J., Madupu, R., Nelson, W., White, O., Peterson, J., Khouri, H., Hance, I., Chris Lee, P., Holtzapple, E., Scanlan, D., Tran, K., Moazzez, A., Utterback, T., Rizzo, M., Lee, K., Kosack, D., Moestl, D., Wedler, H., Lauber, J., Stjepandic, D., Hoheisel, J., Straetz, M., Heim, S., Kiewitz, C., Eisen, J.A., Timmis, K.N., Dusterhoft, A., Tummler, B. and Fraser, C.M. (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol*, **4**, 799-808.

Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. and 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.

Ruis, H. and Schuller, C. (1995) Stress signaling in yeast. *Bioessays*, **17**, 959-965.

Stephanopoulos, G., Aristidou, A.A. and Nielsen, J. (1998) *Metabolic engineering: principles and methodologies*. San Diego: Academic Press.

Stjepandic, D., Weinel, C., Hilbert, H., Koo, H.L., Diehl, F., Nelson, K.E., Tummler, B. and Hoheisel, J.D. (2002) The genome structure of *Pseudomonas putida*: high-resolution mapping and microarray analysis. *Environ Microbiol*, **4**, 819-823.

Wackett, L.P. (2003) *Pseudomonas putida*-a versatile biocatalyst. *Nat Biotechnol*, **21**, 136-138.

Washburn, M.P., Wolters, D. and Yates, J.R., III. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, **19**, 242-247.

Weinel, C., Nelson, K.E. and Tummler, B. (2002) Global features of the *Pseudomonas putida* KT2440 genome sequence. *Environ Microbiol*, **4**, 809-818.

Zhao, Y. and Lin, Y.-H. (2002) Flux distribution and partitioning in *Corynebacterium glutamicum* grown at different specific growth rates. *Process Biochem*, **37**, 775-785.

Appendix C Supplementary data for the developed two-pass approach

Supplementary Data 1 – known sequence

This supplementary document contains three parts of information; Part 1, experimental data; Part 2, a list of peptide candidates that match the m/z value of the precursor ion; Part 3, a list of searched possible peptides that match m/z values of product ions series.

1. Experimental data

Data shown in Table C.1 was derived from Figure 2 of Peng *et al*. (2003). For example, b_5 ion refers to the sequence of HEAAE; hence, the corresponding m/z value was calculated as 538.54. Totally 15 product ions were recalculated to form the MS/MS data seen in Table C.1.

Table C.1 Derived experimental data (known sequence)

Precursor ion (m/z, z)	Product ions series (m/z)
1010.7, +2	538.54, 609.62, 722.78, 779.83, 850.91, 899.07, 964.07, 986.15, 1035.15, 1057.22, 1122.22, 1170.38, 1241.46, 1298.51, 1411.67

2. List of peptide candidates

Table C.2 shown below lists all the matched peptide sequence, its corresponding protein as well as the peptide positions in the protein for the precursor ion at m/z = 1010.7, and z =2. The information presented in the peptides and proteins column of the table is read as; for example, P32795_[568]ALGITFQLPEMDKVDITK[585] refers that the matched peptide is 'ALGITFQLPEMDKVDITK', positioned from 568 to 585 in the protein P32795.

Table C.2 List of matched peptide candidates (known sequence)

	Mass difference (Da)	Peptides and proteins
1	0.0014	P32795_[568]ALGITFQLPEMDKVDITK[585]
2	0.0014	Q02354_[373]YMPVEKLDIDQLQLSVK[389]
3	0.004	P48581_[58]IQLLLSRELFMSYSYR[73]
4	0.0089	P27636_[578]SPVHSLMATRPSSPMRHK[595]
5	0.0098	P47079_[37]ELHQMCLTSMGPCGRNK[53]
6	0.0193	Q03081_[42]IRQSSPLSAVIPAPENVLK[60]
7	0.0198	Q02202_[25]RYFQDNSVLVIPDLLVK[41]
8	0.0198	P20051_[149]EPIHVLNAEEAFLPALKK[166]
9	0.0227	Q04949_[285]LYPLEVFKVNIQEELGK[301]
10	0.0236	P38198_[453]LFQLSFLLINEKTVTPR[469]
11	0.0242	P35209_[358]VSPGIATIAKKPASININPK[377]
12	0.0362	P17883_[144]AFYQHLKPGSLMAETIGR[161]
13	0.0384	P00331_[60]LPLVGGHEGAGVVVGMGENVK[80]
14	0.0384	P00330_[60]LPLVGGHEGAGVVVGMGENVK[80]
15	0.0412	P38811_[2595]QISSRTNVINMLLDSISK[2612]
16	0.0418	P40317_[665]LLFQQLVANDPSMDKATK[682]
17	0.045	Q12333_[492]IVVRIYVCSDSTVPGIIK[509]
18	0.0473	P32559_[389]IIIVVNKSDLVSDDEMTK[406]
19	0.0501	P12686_[156]YAHMVGLLYGIEHKFLK[172]
20	0.0504	P35843_[78]EVAQMLAVVRWFISTLR[94]
21	0.0527	P40537_[167]HKAPCIMTFVSDHNHPK[183]
22	0.0557	P38781_[343]NHQQYMEVCKVNFPPK[358]
23	0.0564	P25573_[265]MEFHLDMFEFFQNKR[279]
24	0.0571	P49090_[285]LHSFAIGLPNAPDLQAARK[303]
25	0.0585	Q12265_[81]SASASRVTAVMPYLCYSR[98]
26	0.0596	P40467_[99]AILPGASTIPASNNPSKPRK[118]
27	0.0614	P38252_[151]TLESSTACAMIPSSLHWK[168]
28	0.0632	Q02486_[10]SFHESSKPLFNLASTLLK[27]
29	0.0639	P36122_[670]FDVDIILDLLVKLISFR[686]
30	0.0655	P47167_[211]ANIITVIEGSTNPGTKYIK[229]
31	0.0656	P40157_[600]TAIPSLGQAIKFITTSADGK[619]
32	0.0657	Q02630_[754]ATVTNTVSYPIQPSATKIK[772]
33	0.0668	P37012_[192]DYVNFLKEIFDFDLIK[207]
34	0.0708	P33310_[237]AIIQGFVGFGMMSFLSWK[254]
35	0.0771	Q04952_[1104]LHYGHPDFLNGIFMTTR[1120]
36	0.0795	P27895_[921]FNDQFEQLINKHNMLK[936]

37	0.0802	Q04373_[260]QQMIKEFWGSEYAVFR[275]
38	0.0821	P38798_[645]DFVIRCIDQVLENIER[660]
39	0.0821	Q02794_[317]NIIENYLLNVAVEAQCR[333]
40	0.0822	P32383_[218]RELMESILLPDNSQFAR[234]
41	0.0853	P38883_[115]QLFNTLISSVAIIIDLMK[132]
42	0.0907	P38069_[132]ELSKCLELSPDEVASLTK[149]
43	0.0914	P36027_[8]NSFRLLLLILSCISTIR[24]
44	0.0999	Q03640_[1470]ASPEANLVLGAISHQRLSR[1488]
45	0.101	P33755_[305]NEMLQIDRQAQEMGLSR[321]
46	0.1021	P53851_[267]NFITGSIDGNCYVWNMK[283]
47	0.1035	P40564_[354]VLSAAWHGSKYEITSTLR[371]
48	0.1042	Q12303_[490]STLGLLLVPSLLILSVFFS[508]
49	0.106	P19414_[409]TIFTVTPGSEQIRATIER[426]
50	0.1066	P32843_[221]YGTIIDIFPPTAANNNVAK[239]
51	0.1071	Q12267_[1121]DVTHTLGMLDDNKMDSVK[1138]
52	0.1091	P47120_[270]HEAAEALGAIASPEVVDVLK[289]
53	0.1091	Q12220_[539]VSPDDRYLAISLLDNTVK[556]
54	0.1095	P80210_[372]VEVEYKVLPGWDQDITK[388]
55	0.1224	P53963_[584]NIEVTVPMHPSEHGTKSR[601]
56	0.123	P53125_[1019]LEDLSRIMDWLDNWGR[1034]
57	0.1262	P40352_[745]IIIFDPDWNPSTDMQAR[761]
58	0.1284	P38257_[159]NPEPIAVDCEYKTQGIGK[176]
59	0.131	Q06156_[505]IENEVENINATNTSVLMK[522]
60	0.131	Q02629_[783]SPNGSTSIPMIENEKISSK[801]
61	0.1341	P13663_[54]QTDLLPESATDIIVSECK[71]
62	0.1468	Q04120_[147]NVNEALRLVEGFQWTDK[163]
63	0.1469	P07806_[300]SVEEAFVRLHDEGVIYR[316]
64	0.1493	Q03825_[75]FLSEADLPLSRINGSASGGK[94]
65	0.1497	P19524_[15]ELGWIGAEVIKNEFNDGK[32]
66	0.1498	P32329_[41]LPFHKLYGDSLENVGSDK[58]
67	0.1498	Q04257_[91]DSHYETLDGKTVVIQWK[107]
68	0.1499	P47014_[512]IPFQHFGATIQISDTTDK[529]
69	0.1505	P53122_[36]FEDQNFQTEFFLNVLK[51]
70	0.1511	P12945_[192]ISDSEKYEHSECLMYK[207]
71	0.1512	P38206_[496]ELFTDSSFFFNFKDFK[511]
72	0.1523	Q04304_[149]NSNLDYTILQPGSLELNK[166]
73	0.1526	P53334_[149]SASEVASDLAQLTDFPVIR[167]
74	0.1555	Q04213_[369]DDFEIILDELQIALDTR[385]
75	0.1556	P13185_[549]SEPEATLATKDTSVPFTPK[567]
76	0.1682	Q06336_[493]GTTLSLQPQSGNMLQSNSR[511]
77	0.1695	Q03231_[80]NNISKTFEDDIFYCPR[95]

78	0.1713	P32912_[203]SLLKECDDIGTANIAQDR[220]
79	0.1719	Q01477_[893]GGEEASDSRTAYILMYQK[910]
80	0.1719	P23643_[174]GNYMLVSQTKFDAESNSK[191]
81	0.1725	P48524_[771]ALYDDFHSKICEYETK[786]
82	0.1929	P32873_[199]IHSEQLASPAASVTYTTSR[217]
83	0.1956	P32567_[211]LDNNGDLLLDTEGYKPNK[228]
84	0.1957	P35194_[1796]YLLGLNHNSDSESESILK[1813]
85	0.1964	P23615_[355]ISVDKNFDANYDLTEFK[371]
86	0.1988	P40084_[193]DDTELEDDLSKWLAQIK[209]
87	0.2148	Q03818_[14]EERSNPQTDSMDDLLIR[30]
88	0.233	P04803_[190]STHVPVGDDQSQHLELTR[207]
89	0.2336	P32336_[83]VPSGFSGTTATSHQEAQWK[101]
90	0.2337	P38800_[547]DAEFHAIFKDSGVSPNER[564]
91	0.236	P35736_[788]SYKVHQAVDGTGEDSIANK[806]
92	0.2367	Q01159_[42]TFPGSQPVSFQHSDVEEK[59]
93	0.2373	P35735_[208]SFNQDYNTVDELPWYK[223]
94	0.2391	Q03707_[198]HLNLLSSDSEIEQDYQK[214]
95	0.2416	Q02206_[403]STTSDIEKTNSLESEHLK[420]
96	0.2452	P47050_[327]SISEYIEIGKDTYDEEK[343]
97	0.2765	Q12753_[231]SGNNWQDSSVSLPAKADSR[249]
98	0.3656	P37838_[631]RTRPDNEDTGDVGESENK[648]
99	0.406	P47005_[641]NLENDSNNNNNNSDTIAR[658]
100	0.4928	P40340_[226]HSRTSNEENDDENDNSR[242]
101	0.6388	P36026_[555]SPHHHHHHHHSSDDSTK[571]
102	0.6524	P38970_[244]NGNGGMNSNATNNVGNGTGNR[264]
103	0.6617	P38928_[457]SSHHSTSTSSYTSSTYTAK[475]
104	0.6625	P06775_[72]RNEDTEQEDINNTNLSK[88]
105	0.7057	P40433_[666]RSNPTSASSSQSELSEQPK[684]
106	0.7353	P53819_[335]ACALNFGAGPRGGAGDEEDR[354]
107	0.7353	P39971_[55]ACALNFGAGPRGGAGDEEDR[74]
108	0.7353	P40889_[271]ACALNFGAGPRGGAGDEEDR[290]
109	0.7353	P53345_[335]ACALNFGAGPRGGAGDEEDR[354]
110	0.7353	P40105_[157]ACALNFGAGPRGGAGDEEDR[176]
111	0.7353	P24088_[271]ACALNFGAGPRGGAGDEEDR[290]
112	0.7353	P40434_[271]ACALNFGAGPRGGAGDEEDR[290]
113	0.7431	Q06489_[315]TTIDNVTETGDDIIVEER[332]
114	0.7462	P25302_[893]NLDNEVVETESSISNNKK[910]
115	0.7487	P25357_[808]GEEFAGELENAERVNDLK[825]
116	0.7493	Q03764_[484]YDCSEDDSFNYLGFCK[499]
117	0.7541	P38873_[1144]QEADEPGSVEYNARLWR[1160]
118	0.7696	P06785_[128]YKTCDDDYTGQGIDQLK[144]

119	0.77	Q07732_[272]DDRIQELEELNSMNDAK[288]
120	0.789	Q03153_[238]NELNLEELYAPENEKSK[254]
121	0.792	Q04487_[54]SDLWSSNKEEELLVSQR[70]
122	0.7939	P38707_[498]EGIDTDAYYWFIDQRK[513]
123	0.795	P38351_[120]RTEYVSNTIAAHDNTSLK[137]
124	0.795	P38351_[121]TEYVSNTIAAHDNTSLKR[138]
125	0.7977	P00358_[53]YAGEVSHDDKHIIVDGHK[70]
126	0.7977	P32789_[252]SSISSFHNSIFGGGKHTEK[270]
127	0.8026	P32288_[241]GDWNGAGCHTNVSTKEMR[258]
128	0.8159	P32639_[188]LMKNITDYETHPDNSNK[204]
129	0.8189	P08018_[628]NQDVHMSEYITERLER[643]
130	0.8221	P00445_[70]THGAPTDEVRHVGDMGNVK[88]
131	0.8245	P53955_[25]MRSEHFNPAYQQQQQK[40]
132	0.8341	P34756_[1625]MSSDSSLCGLASLANEYSK[1643]
133	0.8353	P32381_[232]LARETTALVESYELPDGR[249]
134	0.8379	P38811_[1708]ENSFYIDHLQLNQSIAK[1724]
135	0.8384	P24869_[39]SNANNPALTNFKSTLNSVK[57]
136	0.8433	P14772_[1188]IQYNVDFVFNFRSTNR[1203]
137	0.8434	P32528_[1701]SVFDHQEYLRWINANK[1716]
138	0.8529	P22543_[559]MELLVHLLETKVRPLVK[575]
139	0.8536	P54860_[1]MTAIEDILQITTDPSDTR[18]
140	0.8568	P40858_[145]QAEVGDILNMTDVTTLGSR[163]
141	0.8568	P30665_[473]RLDVDTSTIEQELMQNK[489]
142	0.8592	P36028_[1360]NSISCIPQDPTLFDGTVR[1377]
143	0.8672	P38818_[293]GNSQFWTVSFDRCFLR[308]
144	0.8678	P32432_[438]NAMVNRPHTFNNYSLNK[454]
145	0.8758	Q08960_[346]TGGSAKIDEWTSLLAETLK[364]
146	0.8758	P32873_[664]DDVLQLFDKNQLTETIK[680]
147	0.8769	P09119_[489]IDVDLDMREFYDEMTK[504]
148	0.8787	P12398_[54]IIENAEGSRTTPSVVAFTK[72]
149	0.8787	P39987_[51]IIENAEGSRTTPSVVAFTK[69]
150	0.8788	Q05854_[539]IESQFVETLQLLKNDSR[555]
151	0.8812	Q06488_[381]FELSKPDRSFIPEGELR[397]
152	0.8813	P53550_[118]ILLVQGTESDSWSFPRGK[135]
153	0.8836	P28495_[75]FFDPVNSVIFSVNHLER[91]
154	0.8837	P38704_[357]NCIEATVMQSKERPNDK[373]
155	0.8845	P28272_[217]NGFGGIGGEYVKPTALANVR[236]
156	0.8869	P09436_[81]RFGWDTHGVPIEHIIDK[97]
157	0.8975	P39526_[1875]IISNIFKYPLLQYFMK[1890]
158	0.8976	P53285_[15]LYFLVTFIIYSIIPCR[30]
159	0.899	P49723_[256]EYYSNSLPVEKFGMDLK[272]

160	0.8995	P39946_[200]AITSMDVLFTNYTNSSKK[217]
161	0.902	P40492_[1]MAEKSIFNEPDVDFHLK[17]
162	0.9026	P53167_[317]NLLYCEIRPDDITLER[332]
163	0.905	P53169_[184]YVTTNVQAMDDPHFILR[200]
164	0.9076	P24384_[202]MKNCDGLVHISEMSDQR[218]
165	0.9115	Q02207_[543]RVIGQLFEVGGGWCGQTR[560]
166	0.916	P11745_[25]LTTSDDIKPYLEELAALK[42]
167	0.9184	P32473_[343]ELEDFAFPDTPTIVKAVK[360]
168	0.919	P53955_[181]AILVYLSETASIQDEIVR[198]
169	0.9201	P43560_[656]SIAVSLHQLVKLQLVELK[673]
170	0.9203	Q03631_[631]VIAFYYSVEAYLYQYK[646]
171	0.9215	P38920_[752]DVVEIANLPDLYKVFER[768]
172	0.9217	P25648_[934]LLPINLENNDGSYGLFLK[951]
173	0.9217	P40354_[361]DTVLEKTFLGTSLGQPWK[378]
174	0.9221	P19524_[395]IVSNLNYSQALVAKDSVAK[413]
175	0.9222	P38848_[77]KELLQQIAGSLFSTSIER[94]
176	0.9342	P53247_[29]LAIAIPLLFNLFSRGCGR[46]
177	0.9348	P40989_[1782]IDKFHSIMLFWLKPSR[1797]
178	0.9348	P38631_[1763]IDKFHSIMLFWLKPSR[1778]
179	0.9366	P47014_[690]TLLINNLRHLMLINPDK[706]
180	0.9374	P39523_[822]ELMNELTLVSTELAESIK[839]
181	0.9374	P46673_[681]KLMDEDSVATVIEVIETK[698]
182	0.9428	P18480_[629]MLPTITLDDVYRPAAESK[646]
183	0.943	Q04304_[210]TISLVNGNEPMEKFIQSL[227]
184	0.9436	Q03208_[245]ISKSLDELCGVQLTSTLR[262]
185	0.946	P17442_[499]NVISSTKVQFDPLNVACK[516]
186	0.9461	P20095_[290]VADEMNVVLGKEVGYQIR[307]
187	0.9461	P43606_[414]CIDIDPRSQIIAYGITGK[431]
188	0.9484	P40064_[1275]TTRDTDVVFPVHFLMNK[1291]
189	0.9486	Q12676_[313]KEYGNQPLTFVMAVTHGK[330]
190	0.951	P50077_[1815]IMHSFDGPLSFKIWEGR[1831]
191	0.9587	P25623_[701]YSIKEPIAPIVIHPVWR[717]
192	0.9598	Q02647_[74]GHFVYFYIGPLAFLVFK[90]
193	0.9616	P53140_[284]EFIHPNLYSGLIKVFIK[300]
194	0.9624	Q03213_[234]LAIELLNSISAVSSAYLQK[252]
195	0.9646	Q05568_[35]DEQIQGLLIMKVTELCK[51]
196	0.9654	P16521_[167]MPELIPVLSETMWDTKK[183]
197	0.9668	P36052_[128]NLMSYLKSTLSDNMFQK[144]
198	0.9675	Q03330_[104]VVNNDNTKENMMVLTGLK[121]
199	0.9686	P29539_[130]IPGSSPKPSPSSKPGKSILR[149]
200	0.9688	P36130_[534]VGKPVRLLEGHTDGITSLK[552]

201	0.9717	P38308_[883]VGKPTLRIDSITHNLISR[900]
202	0.9813	Q03103_[268]VTNMYFNYAVVAKALWK[284]
203	0.9894	P38193_[420]MVAIAGITYRENISSPLGK[438]
204	0.9894	P24004_[745]TLLASAVAQQCGLNFISVK[763]
205	0.9967	P53237_[1]MHRMSSTVISLAHFCDK[17]

3. List of possible peptides

Table C.3 below contains the list of searched possible peptides. These peptides are ranked according to the proposed heuristic approach. Column 1 of the table reports m/z value obtained from the experiment; Column 2, predicted m/z value of product ions series; Column 3, ion type; Column 4, predicted charge state; Column 5, predicted peptide sequence. Note that b[1--5] reads as b_5 ion.

Table C.3 List of searched possible peptides (known sequence)

1	P47120_[270]HEAAEALGAIASPEVVDVLK[289]				15
	538.54	538.53774	b[1--5]	[z=1]	HEAAE
	609.62	609.61654	b[1--6]	[z=1]	HEAAEA
	722.78	722.77604	b[1--7]	[z=1]	HEAAEAL
	779.83	779.82804	b[1--8]	[z=1]	HEAAEALG
	850.91	850.90684	b[1--9]	[z=1]	HEAAEALGA
	964.07	964.06634	b[1--10]	[z=1]	HEAAEALGAI
	1035.15	1035.14514	b[1--11]	[z=1]	HEAAEALGAIA
	1122.22	1122.22334	b[1--12]	[z=1]	HEAAEALGAIAS
	899.07	899.07552	y[1--8]	[z=1]	KLVDVVEP
	986.15	986.15372	y[1--9]	[z=1]	KLVDVVEPS
	1057.22	1057.23252	y[1--10]	[z=1]	KLVDVVEPSA
	1170.38	1170.39202	y[1--11]	[z=1]	KLVDVVEPSAI
	1241.46	1241.47082	y[1--12]	[z=1]	KLVDVVEPSAIA
	1298.51	1298.52282	y[1--13]	[z=1]	KLVDVVEPSAIAG
	1411.67	1411.68232	y[1--14]	[z=1]	KLVDVVEPSAIAG L
2	P32873_[199]IHSEQLASPAASVTYTTSR[217]				7
	779.83	779.87144	b[1--7]	[z=1]	IHSEQLA
	964.07	964.06634	b[1--9]	[z=1]	IHSEQLASP
	1035.15	1035.14514	b[1--10]	[z=1]	IHSEQLASPA
	609.62	609.65994	y[1--5]-H2O	[z=1]	RSTTY
	986.15	986.06992	y[1--9]	[z=1]	RSTTYTVSA
	1057.22	1057.14872	y[1--10]	[z=1]	RSTTYTVSAA
	1241.46	1241.34362	y[1--12]	[z=1]	RSTTYTVSAAPS
3	P47014_[512]IPFQHFGATIQISDTTDK[529]				4
	899.07	899.04014	b[1--8]	[z=1]	IPFQHFGA
	1241.46	1241.43554	b[1--11]	[z=1]	IPFQHFGATIQ
	779.83	779.82252	y[1--7]	[z=1]	KDTTDSI
	1122.22	1122.21792	y[1--10]	[z=1]	KDTTDSIQIT
4	P36052_[128]NLMSYLKSTLSDNMFQK[144]				4
	609.62	609.72414	b[1--5]	[z=1]	NLMSY
	722.78	722.88364	b[1--6]	[z=1]	NLMSYL
	850.91	851.05784	b[1--7]	[z=1]	NLMSYLK
	779.83	779.90201	y[1--13]-NH3	[z=2]	KQFMNDSLTSKLY

5	P29539_[130]IPGSSPKPSPSSKPGKSILR[149]				3
	899.07	899.12512	y[1--8]	[z=1]	RLISKGPK
	986.15	986.20332	y[1--9]	[z=1]	RLISKGPKS
	1170.38	1170.39822	y[1--11]	[z=1]	RLISKGPKSSP

Supplementary Data 2 – unknown sequence

This supplementary document contains three parts of information; Part 1, experimental data; Part 2, a list of peptide candidates that match the m/z value of the precursor ion; Part 3, a list of searched possible peptides that match m/z values of product ions series.

1. Experimental data

The experimental MS/MS data (000.30.30.2.dta) of *S. cerevisiae* was retrieved from http://bioinformatics.icmb.utexas.edu/OPD (Prince *et al*., 2004). Mass information shown in Table C.4 was used to verify the proposed heuristic approach.

Table C.4 Derived experimental data (unknown sequence)

Precursor ion ($(M+H)^+$, z)	Product ion series (m/z)
1578.75, +2	242.3, 314.1, 379.3, 389.1, 393.1, 413.3, 416.2, 433.3, 434, 445, 476.2, 496, 497.3, 500, 501.1, 502.9, 506.2, 511.2, 514.7, 528.9, 541, 557.3, 571.8, 576.2, 593.8, 597.1, 599, 600.9, 609.9, 615.5, 616.5, 628.8, 637, 637.7, 640.1, 640.9, 641.7, 645.5, 647.9, 654.8, 658.3, 663.5, 669.3, 671.6, 673, 673.8, 679.6, 687.4, 689.8, 691.1, 693.4, 698.9, 700.3, 703.2, 707, 708.3, 715.8, 718.7, 722.2, 725.2, 727.3, 730.2, 730.9, 736.1, 742.6, 743.6, 744.3, 748.7, 750.5, 751.9, 752.9, 753.7, 754.4, 760.9, 770.5, 771.4, 772.4, 773.4, 788.1, 789, 792.9, 822.4, 823, 827.8, 836, 853.4, 868.4, 871.7, 873.3, 876, 892.4, 893.1, 899.5, 940.2, 994.4, 999.3, 1001.3, 1021.4, 1023.8, 1025.2, 1058.1, 1098.6, 1114.8, 1185.4, 1186.2, 1207.3, 1209.5, 1215.7, 1245.8, 1256.4, 1257.5, 1258.2, 1259.1, 1344.3, 1345.1, 1427.2, 1480.7

2. List of peptide candidates

Table C.5 shown below lists all the matched peptide sequence, its corresponding protein as well as the peptide positions in the protein for the precursor ion at m/z = 1578.75, and z =2. The information presented in the Peptides and proteins column of the table is read as; for example, P38213_[1]MNQNLKNTSWADR[13] refers that the matched peptide is ‘MNQNLKNTSWADR’, positioned from 1 to 13 in the protein P38213.

Table C.5 List of matched peptide candidates (unknown sequence)

	Mass difference (Da)	Peptides and proteins
1	0.0004	P38213_[1]MNQNLKNTSWADR[13]
2	0.002	P54005_[31]SVPECFHFNRER[42]
3	0.0035	P29952_[55]MPSYNHESKESLR[67]
4	0.0058	P38970_[409]TSNLKNGNNDELMK[422]
5	0.0058	P38041_[325]DEIQQQISSKCNK[337]
6	0.0059	P27796_[124]QCSSGLTAVNDIANK[138]
7	0.01	P00447_[132]AIDEQFGSLDELIK[145]
8	0.0106	P23643_[397]SQLLQSITTSGSDLK[411]
9	0.0119	P47019_[117]LEDEMDIDLDGGKK[130]
10	0.0123	P40453_[1056]ISSSDVYVLFYER[1068]
11	0.0129	Q07807_[83]TALSVGTAPPFSTNSK[98]
12	0.013	P14065_[68]DSGVPREEIFVTTK[81]
13	0.0136	P51533_[46]TLTSQSSLLSQEKR[59]
14	0.0155	P24088_[1413]FHPVTDINKESYK[1425]
15	0.0155	P40434_[1376]FHPVTDINKESYK[1388]
16	0.0155	Q03099_[991]FHPVTDINKESYK[1003]
17	0.0155	O13559_[999]FHPVTDINKESYK[1011]
18	0.0155	P38900_[241]FHPVTDINKESYK[253]
19	0.0155	P40105_[1299]FHPVTDINKESYK[1311]
20	0.0155	P53819_[1477]FHPVTDINKESYK[1489]
21	0.0155	P27351_[63]FWQIEDDLEVKR[74]
22	0.0155	P38909_[93]ENKHELSPSYFVK[105]
23	0.0155	P40889_[1376]FHPVTDINKESYK[1388]
24	0.0155	P53345_[1477]FHPVTDINKESYK[1489]
25	0.0156	P40573_[151]IEQLNKENEFWK[162]
26	0.0157	P52893_[64]HSSSWIVAQNHRR[76]
27	0.0159	P32074_[547]DRATIALEFIDSAR[560]
28	0.016	Q02197_[624]LERATNFIETEVR[636]
29	0.0192	P18888_[65]SLTYAQQQLNKQR[77]
30	0.0217	P25335_[241]QPGHTDWAVIQLGR[254]
31	0.0223	P38081_[432]NFEHWRFEDGIK[443]
32	0.024	P40477_[1004]SGQPNHGVQGDGIALK[1019]
33	0.0254	P32898_[188]GQISNANYYFWSK[200]
34	0.0267	P34250_[120]VLNISSSTGQNSKSR[134]
35	0.0272	Q03660_[200]GNNFEEQLLTREK[212]
36	0.0272	P22148_[293]EYTEGVNGQPSIRK[306]
37	0.0273	P40061_[402]DNSFQIEREQALK[414]
38	0.0304	P17255_[408]EVSKSYPISEGPER[421]
39	0.0304	P47096_[149]EAILDFENDVEKR[161]
40	0.0304	P38144_[558]IQAIDDYNAPDSKK[571]

41	0.0305	P39717_[21]VAVSPFSSALEGEER[35]
42	0.031	P43634_[232]ETFLSAFFGDTNTK[245]
43	0.0328	P10834_[21]VLGSVESGNSATISEK[36]
44	0.0335	P09119_[33]LQFTDVTPESSPEK[46]
45	0.0344	P34111_[594]TVRGDVDLMVESEK[607]
46	0.0363	P32849_[940]CLFEYIEFQNSK[951]
47	0.0368	Q04693_[187]VSPISYMEIDPNGR[200]
48	0.0369	P29539_[1728]DEGFLKSMEHAVSK[1741]
49	0.043	Q12222_[4]VNNVFGSNPNRMTK[17]
50	0.0431	P39956_[729]KSVHSGEKPHSCPK[742]
51	0.0438	P40527_[56]HNTVGDRESFEMR[68]
52	0.0493	P47027_[629]NNSSLSEHSMKDTK[642]
53	0.0499	P12904_[264]RSDDFEGVYTCTK[276]
54	0.0529	P51979_[963]CLFYESSSDGEVGK[976]
55	0.0533	P38197_[105]VPNLYSVETIDSLK[118]
56	0.0534	Q05506_[413]GTVVFLDNILEETK[426]
57	0.0534	P04050_[258]GEDDLTFKLADILK[271]
58	0.0539	P38742_[468]SIASKISSLSENTLK[482]
59	0.0564	P43535_[737]DYILQSADAAGVVKK[751]
60	0.0564	P53227_[44]DNVLLASEFKINSK[57]
61	0.0565	P53691_[218]AIETVKNIGTEQFK[231]
62	0.0582	P43561_[359]YAFFNNITYVTPK[371]
63	0.0588	P38870_[647]YSLPPQTIQDLFR[659]
64	0.0589	P33330_[102]IAPAGYLVTGSWSQK[116]
65	0.0589	Q03722_[208]QTIWNTVTSTIWK[220]
66	0.059	P40345_[265]VFQNLGVIGYEPNK[278]
67	0.0595	P08536_[394]QGFSIVLGNSLTVSR[408]
68	0.0595	P47029_[431]GPKLPNLPNDANLSK[445]
69	0.0596	P80210_[295]LQTIGAEFGVTTGRK[309]
70	0.0619	P21827_[153]IPRESFPPLAEGHK[166]
71	0.062	Q12514_[157]HEFHIANLENILK[169]
72	0.062	P53950_[1048]LETPLKFQGGAFNR[1061]
73	0.0638	P38083_[253]VPMGCDVSLSHYGR[266]
74	0.0645	P43555_[399]RHGQDGPQVDEIAR[412]
75	0.0656	Q12451_[488]TPVGVHTGSALQRVR[502]
76	0.0668	P22470_[269]WSRLENSCPLCR[280]
77	0.0686	P50111_[225]HGNASLIRRPSTLR[238]
78	0.0702	P53951_[296]TDNVTNSSRSIAANK[310]
79	0.0706	P34237_[166]TQEINSTWEEKGR[178]
80	0.0713	Q12019_[3570]QHFYEDPNLEASK[3582]
81	0.0731	P39521_[174]TEGIRNSEDTSIQK[187]
82	0.0731	P32829_[421]SGINGTSISDRDVEK[435]
83	0.0737	P38272_[170]DHYSDEISKLNEK[182]
84	0.0737	Q02785_[1494]DTNIFQTVPGDENK[1507]
85	0.0795	P38687_[433]YMEALALYVDAYR[445]

86	0.0802	P15442_[766]DLKPMNIFIDESR[778]
87	0.0803	P38866_[58]VMYDHLETNISKK[70]
88	0.0828	P53210_[107]FSGQCFTISKQFK[119]
89	0.0828	P21268_[520]FNENCEKWLLPK[531]
90	0.0832	P20448_[342]AQQVCLFATDVVAR[355]
91	0.0833	P21951_[1652]LSQYSNIPICNLR[1664]
92	0.0833	P15801_[606]ERIQSQFVVPSCK[618]
93	0.0833	P40014_[14]MDGFQKDVAQVLAR[27]
94	0.0834	Q10740_[163]CTALQWLNSKQTK[175]
95	0.0858	P35176_[74]DPKMGYLNSIFHR[86]
96	0.0902	P07265_[181]QVDLNWENEDCR[192]
97	0.0902	P38158_[181]QVDLNWENEDCR[192]
98	0.0902	P53341_[181]QVDLNWENEDCR[192]
99	0.0927	P32912_[207]ECDDIGTANIAQDR[220]
100	0.0991	P38297_[635]VSLSITDLFAPTWK[648]
101	0.0998	P53276_[360]TFQLLKSAVINSEK[373]
102	0.101	P53728_[81]TVMTFCQYVDSVK[93]
103	0.1017	P43554_[398]KSEENEMIKPMNK[410]
104	0.1023	P53861_[163]KQTIVVDHTVYFK[175]
105	0.1023	P00950_[17]NLFTGWVDVKLSAK[30]
106	0.1023	P40317_[582]VNLNTSLLWFDKK[594]
107	0.1023	P53861_[164]QTIVVDHTVYFKK[176]
108	0.1028	P13099_[189]AGVTFTRLLTETLR[202]
109	0.1029	Q06406_[60]TVNVKLASGLLYSGR[74]
110	0.104	Q03210_[320]AMVECSLAYRYSK[332]
111	0.1049	P33306_[4]NSHHHRSSSVNSTK[17]
112	0.1082	Q08217_[102]APVIAYPPSLRHTR[115]
113	0.1134	Q07084_[90]SNGAGSGANLSVNSNTK[106]
114	0.1134	P04821_[1543]SGNTKGSTHASSASGTK[1559]
115	0.1165	P32499_[328]KNDENSTSNSKPEK[341]
116	0.1166	P53125_[281]SNSANVSSPESEKNK[295]
117	0.1226	P38151_[29]GEDTSEEQLEAEIK[42]
118	0.1226	P39993_[1004]GDEEPTEEEIKSSK[1017]
119	0.1236	P38811_[2379]MLAFEIRGEPSLSK[2392]
120	0.1237	P25694_[106]LGDLVTIHPCPDIK[119]
121	0.1237	Q12019_[1090]TSMIKYLADITGHK[1103]
122	0.1267	P53212_[49]INNKFANQIAMSVK[62]
123	0.1267	P27636_[256]NMYKRPTADQLLK[268]
124	0.1292	P23292_[214]VHLIDFGMAKQYR[226]
125	0.142	P39985_[925]DMNKDIELMDLLK[937]
126	0.1426	P47054_[1325]SNFILEVFGTIIPK[1338]
127	0.1426	P24482_[402]FVILGANLFLDDLK[415]
128	0.1432	P46784_[39]NLYVIKALQSLTSK[52]
129	0.1432	Q08745_[39]NLYVIKALQSLTSK[52]
130	0.1445	P09064_[229]YILEYVTCKTCK[240]

131	0.1499	P53243_[8]YICSFCLKPFSR[19]
132	0.153	P32639_[1139]MWPTNCPLRQFK[1150]
133	0.1568	P39705_[478]SRSNLSQENDNEGK[491]
134	0.1599	P36224_[90]LSNDEEDESRQQK[102]
135	0.163	P32445_[121]KLEDAEGQENAASSE[135]
136	0.1658	P35999_[745]CIADVLECPMLEK[757]
137	0.1664	P43612_[362]HVDISLLMDFFLK[374]
138	0.1669	P53737_[1]MVQPAPLITNAPTPK[15]
139	0.1671	P25389_[965]NLVFNITNMIITGK[978]
140	0.1694	P16151_[295]MYHSAILVDFLLR[307]
141	0.1757	P14904_[249]SPLFGKHCIHLLR[261]
142	0.1891	Q06053_[28]GIAHIKPEYIVPLK[41]
143	0.2154	P40989_[1691]MLIGVVTCIQCQR[1703]
144	0.2154	P38631_[1672]MLIGVVTCIQCQR[1684]
145	0.2507	P32599_[629]LIITFIASLMTLNK[642]
146	0.2592	P47094_[107]LVFLSKPFRLAMR[119]
147	0.3405	P22023_[3]LLALVLLLLCAPLR[16]
148	0.7119	P54072_[486]LIVIPIIGVLWVNK[499]
149	0.751	P47821_[76]IYCYFLIMKLGR[87]
150	0.7756	P47155_[15]VAFLFTIAFFCLK[27]
151	0.7926	P21954_[136]LVPRWEKPIIIGR[148]
152	0.8153	P39946_[384]IWEIPLPTLMAHR[396]
153	0.8207	P43610_[542]INPTLLQMDKLYK[554]
154	0.8384	Q01846_[197]IHKSLSLKPNALQK[210]
155	0.8385	Q06287_[119]GILIEVNPTVRIPR[132]
156	0.8391	P12686_[57]TDAKLKPFIYRPK[69]
157	0.8396	P38954_[427]WIFKIVNDGFIPK[439]
158	0.8415	P40468_[609]TLTKLLQLYLNTR[621]
159	0.8587	P42945_[157]LPPLFNCLSNFVR[169]
160	0.8611	P53204_[359]GMSVQYLLPNSVIR[372]
161	0.8635	P06779_[162]VNKVSSLQSLCITK[175]
162	0.8641	P12294_[295]NSILVEKWMDTLK[307]
163	0.8641	P38863_[715]AKLCQLDPVLYEK[727]
164	0.8672	P48164_[125]VLEDMVFPTEIVGK[138]
165	0.8682	P25558_[793]DSPSGDNSNVTKETK[807]
166	0.8682	P54791_[196]NEDSGEVDRESITK[209]
167	0.8682	Q03661_[1101]DQDSTAEKNVEGSAK[1115]
168	0.8712	P11927_[71]NINSDSDRSNDTIK[84]
169	0.8731	P33332_[16]ETSHDENTSFFHK[28]
170	0.8733	P04650_[18]QNRPLPQWIRLR[29]
171	0.8774	Q07505_[124]IGSTGMCLGGHLAFR[138]
172	0.88	P35182_[108]LVGNSGCTAAVCVLR[122]
173	0.8812	P53318_[217]GWMYNAYGVVASMK[230]
174	0.8825	P38873_[894]KELVVYFSHIVSR[906]
175	0.8825	P32917_[156]VAPFGYPIQRTSIK[169]

176	0.8831	P53268_[44]GVKDIFSFFFLTR[56]
177	0.8849	Q00955_[325]HQKIIEEAPVTIAK[338]
178	0.8849	P00635_[308]SVGSNLFNASVKLLK[322]
179	0.8855	P22149_[121]IFYPQGIELVIER[133]
180	0.8856	P40032_[168]ISFILYLPDPDRK[180]
181	0.8879	P32893_[495]KIAEGIILLSNDYK[508]
182	0.888	P51533_[358]GLDSATALEFIKALK[372]
183	0.8886	P20049_[110]LPEIEAFEKYLPK[122]
184	0.9006	P53745_[43]EDDRSGNVHCFSR[55]
185	0.9025	P40020_[844]LSHCNEILGMCDK[856]
186	0.9045	P13298_[11]NFLELAIECQALR[23]
187	0.905	P21560_[112]AGPVAGSYYYKICK[125]
188	0.9051	Q99258_[1]MFTPIDQAIEHFK[13]
189	0.9055	P46949_[435]GSILLTSDEEEEEK[448]
190	0.9069	P89105_[881]IQLGETTMKSALER[894]
191	0.907	P53960_[328]LARTASEELMNTLK[341]
192	0.907	P32642_[39]TMSQVLEAVSEKVR[52]
193	0.9075	P34243_[59]LYMELGPNLAVNDK[72]
194	0.9076	Q04660_[250]HEEVMPLTAVPEPK[263]
195	0.9084	P29547_[238]EEAKPAATETETSSK[252]
196	0.9085	P53552_[377]DKISEETNADIESK[390]
197	0.9099	P29465_[431]CLSEIIKVGEVDSK[444]
198	0.9103	Q06010_[131]FPDENEYSSYLSK[143]
199	0.911	P39105_[45]EASGLSDNETEWLK[58]
200	0.9116	P40036_[47]ERDGSTEETLNSLK[60]
201	0.9135	P36002_[150]DPDLGFYLHDGDSK[163]
202	0.9165	Q06245_[855]EDFNHDNFINSVK[867]
203	0.9191	Q02208_[159]NQHISLLQLARQR[171]
204	0.9204	P33334_[1605]RFTLWWSPTINR[1616]
205	0.9227	P32457_[94]RQINGYVGFANLPK[107]
206	0.9228	P17883_[1075]LFAHSFILSNGRSK[1088]
207	0.9229	Q10740_[10]HSSSIYLPTLRFR[22]
208	0.9253	P26793_[5]GLNAIISEHVPSAIR[19]
209	0.9253	P00812_[101]LVYNSVSKVVQANR[114]
210	0.9255	Q12117_[285]APVASPRPAATPNLSK[300]
211	0.9258	Q12676_[229]LPLNGEYQIFNLR[241]
212	0.9258	P38150_[461]ISRIYPELYHTGK[473]
213	0.9259	P25648_[790]NFPFVLKVDNDLR[802]
214	0.926	P53043_[213]YVAAIISHADTLFR[226]
215	0.9264	P53327_[302]TNENMLICAPTGAGK[316]
216	0.9282	P38737_[422]QLQYETLGDILKR[434]
217	0.9282	P40531_[179]IQQDTLIQTKFNK[191]
218	0.9283	P38110_[2203]VLLQYNRDSEVLK[2215]
219	0.9283	Q03435_[36]LEYGVLLERLESR[48]
220	0.9284	P18412_[289]TNAASQAKTPLIYAK[303]

221	0.9284	P40522_[84]EFEKLVTAAVQSVR[97]
222	0.9288	Q07648_[100]GHIAKELYEEFLK[112]
223	0.929	P42883_[227]ATDYVLADPVKAWK[240]
224	0.929	P43534_[227]ATDYVLADPVKAWK[240]
225	0.929	P47183_[227]ATDYVLADPVKAWK[240]
226	0.9293	P15891_[122]DEDDLDENELLMK[134]
227	0.9312	Q06156_[922]ENELLFGEKSILGK[935]
228	0.9344	P26448_[24]YLILSEGLPISEDK[37]
229	0.9478	P53753_[956]LWGATIGDQSMELR[969]
230	0.9484	P25335_[76]HNEMEYDWVIIK[87]
231	0.9502	P00830_[94]TIAMDGTEGLVRGEK[108]
232	0.9503	P53628_[286]VGSENVECTISILR[299]
233	0.9503	P27344_[69]LSLNSIEECVEKR[81]
234	0.9509	P40010_[334]ILDSPGICFPSENK[347]
235	0.9514	P38958_[87]IQEEFGLDLEEEK[99]
236	0.9544	Q00723_[140]NDDDENDLIKLFK[152]
237	0.9573	Q03655_[203]SIPVGYSAADNTDLR[217]
238	0.9574	P08539_[163]AKAAFDEDGNISNVK[177]
239	0.9574	P53290_[126]VDGLSEDFKVDAQR[139]
240	0.9574	Q02825_[275]ELLDAQSDREFQK[287]
241	0.9574	Q04693_[449]NDFSNVLTSKDPNK[462]
242	0.9574	P19146_[83]NTEGVIFVIDSNDR[96]
243	0.958	P35187_[1368]AAASSNGIAQSTGTKSK[1384]
244	0.9599	P25648_[504]QFDHYESNQLVAK[516]
245	0.9605	P34218_[672]NTNNDRLIYQAEK[684]
246	0.9629	Q03834_[277]YQWLVDERDAQR[288]
247	0.963	P40359_[132]DDFHDPIHELRGK[144]
248	0.9635	P38751_[44]EHVTTNTVAGHVASR[58]
249	0.9647	Q04673_[398]FRSEDCFSCQSR[409]
250	0.9661	P41921_[113]DAYVHRLNGIYQK[125]
251	0.9662	P33314_[549]NLDSKHVFNSLFR[561]
252	0.9663	P33329_[318]REYLSAYPTLAHR[330]
253	0.9674	P38181_[992]CGSFCSASDILGFR[1005]
254	0.9686	P47161_[134]IYLNSLQQENRAK[146]
255	0.9686	P27801_[855]VILNSNDYNLRQK[867]
256	0.9687	P47160_[105]FIDDTRNSINLIR[117]
257	0.9687	P46672_[362]GESFKVASIANAQVR[376]
258	0.9692	P40988_[406]EVYFRIVQHEEK[417]
259	0.9693	P53165_[264]SKPYDVLLADYHR[276]
260	0.9716	P40024_[95]VLIQDSGLELNYGR[108]
261	0.9717	P38805_[142]RNFTDIVIINEDK[154]
262	0.9717	P10566_[158]IQLNTSASVWQTTK[171]
263	0.9718	Q07807_[555]FIQRELATSPASEK[568]
264	0.9723	P38732_[437]IYSNNKEFSLSFK[449]
265	0.9748	Q02773_[296]IAIELFNTTTNDPK[309]

266	0.9748	P38999_[293]EDLIASIDSKATWK[306]
267	0.9748	P09457_[138]GTVTSAEPLDPKSFK[152]
268	0.9788	P32590_[654]ANQETSKMNDIAEK[667]
269	0.9799	P38822_[156]DYDEACSTMEMAR[168]
270	0.9843	P40483_[368]EECTWNRLDTVR[379]
271	0.9856	P53048_[354]SGSFFNCFKGVNGR[367]
272	0.9857	P40029_[170]NPQGYACPTHYLR[182]
273	0.9907	P25386_[801]NVRDSLDEMTQLR[813]
274	0.9907	P38342_[42]SEARLLDTCNEIR[54]
275	0.9937	P25045_[402]LMDSNNDAVQTLQK[415]
276	0.9946	P39102_[56]EITAAYEILSDPEK[69]
277	0.9953	P33892_[363]KISNTDTTLEDLTK[376]
278	0.9977	P00924_[88]AVDDFLISLDGTANK[102]
279	0.9977	P00925_[88]AVDDFLLSLDGTANK[102]
280	0.9978	P38902_[75]IQTTEGYDPKDALK[88]
281	0.9978	Q07505_[95]IKKPLESYDEDNK[107]
282	0.9978	P29509_[255]IVAGQVDTDEAGYIK[269]
283	0.9996	P40453_[1032]SKWYYFDDEVVK[1043]

3. List of possible peptides

Table C.6 below contains the list of searched possible peptides. These peptides are ranked according to the proposed heuristic approach. Column 1 of the table reports m/z value obtained from the experiment; Column 2, predicted m/z value of product ions series; Column 3, ion type; Column 4, predicted charge state; Column 5, predicted peptide sequence. Note that b[1--5] reads as b_5 ion.

Table C.6 List of searched peptides (unknown sequence)

1	P32445_[121]KLEDAEGQENAASSE[135]			37
	242.3	242.3416	b[1--2]	[z=1]	KL
	501.1	501.0231	b[1--9]	[z=2]	KLEDAEGQE
	541	540.594	b[1--5]-NH3	[z=1]	KLEDA
	557.3	557.6245	b[1--5]	[z=1]	KLEDA
	593.8	593.6145	b[1--11]	[z=2]	KLEDAEGQENA
	628.8	629.1539	b[1--12]	[z=2]	KLEDAEGQENAA
	663.5	663.6854	b[1--13]-H2O	[z=2]	KLEDAEGQENAAS
	669.3	668.7248	b[1--6]-H2O	[z=1]	KLEDAE
	669.3	669.7095	b[1--6]-NH3	[z=1]	KLEDAE
	673	672.693	b[1--13]	[z=2]	KLEDAEGQENAAS
	687.4	686.74	b[1--6]	[z=1]	KLEDAE
	707	707.2245	b[1--14]-H2O	[z=2]	KLEDAEGQENAASS
	725.2	725.7768	b[1--7]-H2O	[z=1]	KLEDAEG
	727.3	726.7615	b[1--7]-NH3	[z=1]	KLEDAEG
	743.6	743.792	b[1--7]	[z=1]	KLEDAEG
	744.3	743.792	b[1--7]	[z=1]	KLEDAEG
	771.4	771.7822	b[1--15]-H2O	[z=2]	KLEDAEGQENAASSE
	772.4	772.2746	b[1--15]-NH3	[z=2]	KLEDAEGQENAASSE
	853.4	853.9076	b[1--8]-H2O	[z=1]	KLEDAEGQ
	871.7	871.9228	b[1--8]	[z=1]	KLEDAEGQ
	1001.3	1001.038	b[1--9]	[z=1]	KLEDAEGQE
	1098.6	1098.112	b[1--10]-NH3	[z=1]	KLEDAEGQEN
	1114.8	1115.142	b[1--10]	[z=1]	KLEDAEGQEN
	1186.2	1186.221	b[1--11]	[z=1]	KLEDAEGQENA
	1257.5	1257.3	b[1--12]	[z=1]	KLEDAEGQENAA
	1344.3	1344.378	b[1--13]	[z=1]	KLEDAEGQENAAS
	1345.1	1344.378	b[1--13]	[z=1]	KLEDAEGQENAAS
	393.1	393.3739	y[1--4]	[z=1]	ESSA
	502.9	502.9739	y[1--10]-NH3	[z=2]	ESSAANEQGE
	511.2	511.4892	y[1--10]	[z=2]	ESSAANEQGE
	669.3	669.1306	y[1--13]	[z=2]	ESSAANEQGEADE
	689.8	689.6568	y[1--7]-H2O	[z=1]	ESSAANE
	691.1	690.6416	y[1--7]-NH3	[z=1]	ESSAANE
	707	707.6721	y[1--7]	[z=1]	ESSAANE
	708.3	707.6721	y[1--7]	[z=1]	ESSAANE
	836	835.8029	y[1--8]	[z=1]	ESSAANEQ
	876	875.8244	y[1--9]-NH3	[z=1]	ESSAANEQG
	892.4	892.8549	y[1--9]	[z=1]	ESSAANEQG
	893.1	892.8549	y[1--9]	[z=1]	ESSAANEQG

	1021.4	1021.97	y[1--10]	[z=1]	ESSAANEQGE
2	P23643_[397]SQLLQSITTSGSDLK[411]				26
	593.8	593.6579	b[1--12]-H2O	[z=2]	SQLLQSITTSGS
	593.8	594.1502	b[1--12]-NH3	[z=2]	SQLLQSITTSGS
	640.1	639.7297	b[1--6]-H2O	[z=1]	SQLLQS
	640.1	640.7144	b[1--6]-NH3	[z=1]	SQLLQS
	640.9	640.7144	b[1--6]-NH3	[z=1]	SQLLQS
	658.3	657.7449	b[1--6]	[z=1]	SQLLQS
	708.3	708.2743	b[1--14]-NH3	[z=2]	SQLLQSITTSGSDL
	752.9	752.8892	b[1--7]-H2O	[z=1]	SQLLQSI
	753.7	753.8739	b[1--7]-NH3	[z=1]	SQLLQSI
	754.4	753.8739	b[1--7]-NH3	[z=1]	SQLLQSI
	770.5	770.9044	b[1--7]	[z=1]	SQLLQSI
	771.4	770.9044	b[1--7]	[z=1]	SQLLQSI
	772.4	772.3614	b[1--15]-NH3	[z=2]	SQLLQSITTSGSDLK
	853.4	853.9943	b[1--8]-H2O	[z=1]	SQLLQSIT
	871.7	872.0095	b[1--8]	[z=1]	SQLLQSIT
	1098.6	1099.23	b[1--11]-H2O	[z=1]	SQLLQSITTSG
	1186.2	1186.308	b[1--12]-H2O	[z=1]	SQLLQSITTSGS
	445	444.5084	y[1--4]-H2O	[z=1]	KLDS
	445	445.4932	y[1--4]-NH3	[z=1]	KLDS
	496	496.0472	y[1--10]-H2O	[z=2]	KLDSGSTTIS
	501.1	501.5604	y[1--5]-H2O	[z=1]	KLDSG
	502.9	502.5452	y[1--5]-NH3	[z=1]	KLDSG
	616.5	616.6924	y[1--12]-H2O	[z=2]	KLDSGSTTISQL
	673	673.2721	y[1--13]-H2O	[z=2]	KLDSGSTTISQLL
	673.8	673.7645	y[1--13]-NH3	[z=2]	KLDSGSTTISQLL
	689.8	689.7437	y[1--7]-H2O	[z=1]	KLDSGST
	691.1	690.7285	y[1--7]-NH3	[z=1]	KLDSGST
	707	707.759	y[1--7]	[z=1]	KLDSGST
	708.3	707.759	y[1--7]	[z=1]	KLDSGST
	1345.1	1345.536	y[1--13]-H2O	[z=1]	KLDSGSTTISQLL
3	P32074_[547]DRATIALEFIDSAR[560]				24
	445	444.468	b[1--4]	[z=1]	DRAT
	501.1	501.0677	b[1--9]-NH3	[z=2]	DRATIALEF
	541	540.597	b[1--5]-NH3	[z=1]	DRATI
	557.3	557.6275	b[1--5]	[z=1]	DRATI
	557.3	557.1551	b[1--10]-H2O	[z=2]	DRATIALEFI
	557.3	557.6474	b[1--10]-NH3	[z=2]	DRATIALEFI
	609.9	610.6911	b[1--6]-H2O	[z=1]	DRATIA
	615.5	615.1917	b[1--11]-NH3	[z=2]	DRATIALEFID

628.8	628.7063	b[1--6]	[z=1]	DRATIA
658.3	658.2385	b[1--12]-H2O	[z=2]	DRATIALEFIDS
693.4	693.7779	b[1--13]-H2O	[z=2]	DRATIALEFIDSA
725.2	724.8353	b[1--7]-NH3	[z=1]	DRATIAL
742.6	741.8658	b[1--7]	[z=1]	DRATIAL
772.4	772.364	b[1--14]-NH3	[z=2]	DRATIALEFIDSAR
853.4	852.9661	b[1--8]-H2O	[z=1]	DRATIALE
853.4	853.9508	b[1--8]-NH3	[z=1]	DRATIALE
871.7	870.9813	b[1--8]	[z=1]	DRATIALE
1001.3	1001.127	b[1--9]-NH3	[z=1]	DRATIALEF
1114.8	1114.287	b[1--10]-NH3	[z=1]	DRATIALEFI
1245.8	1246.406	b[1--11]	[z=1]	DRATIALEFID
476.2	476.0377	y[1--8]	[z=2]	RASDIFEL
502.9	502.5695	y[1--9]-H2O	[z=2]	RASDIFELA
502.9	503.0619	y[1--9]-NH3	[z=2]	RASDIFELA
511.2	511.5771	y[1--9]	[z=2]	RASDIFELA
609.9	609.7018	y[1--11]-H2O	[z=2]	RASDIFELAIT
609.9	610.1942	y[1--11]-NH3	[z=2]	RASDIFELAIT
645.5	645.2412	y[1--12]-H2O	[z=2]	RASDIFELAITA
645.5	645.7336	y[1--12]-NH3	[z=2]	RASDIFELAITA
691.1	690.7772	y[1--6]-H2O	[z=1]	RASDIF
691.1	691.762	y[1--6]-NH3	[z=1]	RASDIF
708.3	708.7925	y[1--6]	[z=1]	RASDIF
1021.4	1022.146	y[1--9]	[z=1]	RASDIFELA

4 P25558_[793]DSPSGDNSNVTKETK[807] 24

445	444.4216	b[1--5]	[z=1]	DSPSG
528.9	529.0133	b[1--11]-H2O	[z=2]	DSPSGDNSNVT
541	541.495	b[1--6]-H2O	[z=1]	DSPSGD
593.8	593.5928	b[1--12]-NH3	[z=2]	DSPSGDNSNVTK
654.8	655.5989	b[1--7]-H2O	[z=1]	DSPSGDN
658.3	658.1505	b[1--13]-NH3	[z=2]	DSPSGDNSNVTKE
673	673.6141	b[1--7]	[z=1]	DSPSGDN
673.8	673.6141	b[1--7]	[z=1]	DSPSGDN
708.3	708.2107	b[1--14]-H2O	[z=2]	DSPSGDNSNVTKET
742.6	742.6771	b[1--8]-H2O	[z=1]	DSPSGDNS
743.6	743.6618	b[1--8]-NH3	[z=1]	DSPSGDNS
744.3	743.6618	b[1--8]-NH3	[z=1]	DSPSGDNS
760.9	760.6923	b[1--8]	[z=1]	DSPSGDNS
772.4	772.2978	b[1--15]-H2O	[z=2]	DSPSGDNSNVTKETK
772.4	772.7902	b[1--15]-NH3	[z=2]	DSPSGDNSNVTKETK
1058.1	1058.003	b[1--11]-NH3	[z=1]	DSPSGDNSNVT
1185.4	1185.193	b[1--12]-H2O	[z=1]	DSPSGDNSNVTK
1185.4	1186.178	b[1--12]-NH3	[z=1]	DSPSGDNSNVTK

	1186.2	1186.178	b[1--12]-NH3	[z=1]	DSPSGDNSNVTK
	445	445.0023	y[1--8]-H2O	[z=2]	KTEKTVNS
	502.9	502.5467	y[1--9]-NH3	[z=2]	KTEKTVNSN
	506.2	505.5922	y[1--4]	[z=1]	KTEK
	511.2	511.0619	y[1--9]	[z=2]	KTEKTVNSN
	597.1	597.1322	y[1--11]	[z=2]	KTEKTVNSNDG
	640.9	640.6713	y[1--12]	[z=2]	KTEKTVNSNDGS
	687.4	687.8146	y[1--6]-H2O	[z=1]	KTEKTV
	1021.4	1021.116	y[1--9]	[z=1]	KTEKTVNSN
5	P35187_[1368]AAASSNGIAQSTGTKSK[1384]			23	
	389.1	388.4007	b[1--5]	[z=1]	AAASS
	502.9	502.5046	b[1--6]	[z=1]	AAASSN
	541	541.5414	b[1--7]-H2O	[z=1]	AAASSNG
	600.9	600.6294	b[1--14]-H2O	[z=2]	AAASSNGIAQSTGT
	600.9	601.1218	b[1--14]-NH3	[z=2]	AAASSNGIAQSTGT
	609.9	609.637	b[1--14]	[z=2]	AAASSNGIAQSTGT
	654.8	654.7009	b[1--8]-H2O	[z=1]	AAASSNGI
	673	672.7161	b[1--8]	[z=1]	AAASSNGI
	673.8	673.7241	b[1--15]	[z=2]	AAASSNGIAQSTGTK
	708.3	708.2556	b[1--16]-H2O	[z=2]	AAASSNGIAQSTGTKS
	725.2	725.7797	b[1--9]-H2O	[z=1]	AAASSNGIA
	727.3	726.7644	b[1--9]-NH3	[z=1]	AAASSNGIA
	743.6	743.7949	b[1--9]	[z=1]	AAASSNGIA
	744.3	743.7949	b[1--9]	[z=1]	AAASSNGIA
	772.4	772.3427	b[1--17]-H2O	[z=2]	AAASSNGIAQSTGTKSK
	853.4	853.9105	b[1--10]-H2O	[z=1]	AAASSNGIAQ
	871.7	871.9257	b[1--10]	[z=1]	AAASSNGIAQ
	940.2	940.9887	b[1--11]-H2O	[z=1]	AAASSNGIAQS
	1098.6	1099.146	b[1--13]-H2O	[z=1]	AAASSNGIAQSTG
	445	445.5396	y[1--4]-H2O	[z=1]	KSKT
	502.9	502.5916	y[1--5]-H2O	[z=1]	KSKTG
	502.9	503.5764	y[1--5]-NH3	[z=1]	KSKTG
	502.9	502.5684	y[1--10]-NH3	[z=2]	KSKTGTSQAI
	511.2	511.0836	y[1--10]	[z=2]	KSKTGTSQAI
	640.1	640.2007	y[1--13]	[z=2]	KSKTGTSQAIGNS
	691.1	690.7749	y[1--7]-H2O	[z=1]	KSKTGTS
	691.1	691.7597	y[1--7]-NH3	[z=1]	KSKTGTS
	708.3	708.7902	y[1--7]	[z=1]	KSKTGTS
	1021.4	1021.159	y[1--10]	[z=1]	KSKTGTSQAI

References

Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. and 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.

Prince, J.T., Carlson, M.W., Wang, R., Lu, P. and Marcotte, E.M. (2004) The need for a public proteomics repository. *Nat Biotechnol*, **22**, 471-472.

Appendix D Supplementary data for phosphopeptide identification

A: Deduction of peptide sequences using mass information of precursor ions

Given the m/z value and z of a precursor ion, the peptide sequences deduced from the pooled peptide database that match the experimental data are reported. The output format includes the protein name, the start and stop position (embraced by brackets) of the deduced peptide sequence, and the number of the phosphate group (i.e., num_of_phos) in the deduced peptide sequence.

For precursor ion: m/z = 798.31, z = 2
The possible peptides at m/z = 798.31 are:
protein3_[1]KDSDDEEEVVHVD[13] (z=2) (num_of_phos=1)

For precursor ion: m/z = 861.6, z = 3
The possible peptides at m/z = 861.6 are:
protein2_[1]LCDFGVSGQLIDSMANSFVGTR[22] (z=3) (num_of_phos=2)

For precursor ion: m/z = 1046.52, z = 2
The possible peptides at m/z = 1046.52 are:
protein4_[1]DNRSQVETEDLILKPGVV[18] (z=2) (num_of_phos=1)

For precursor ion: m/z = 1101.48, z = 2
The possible peptides at m/z = 1101.48 are:
protein5_[1]EKKEFLEPDSWETLDQQ[17] (z=2) (num_of_phos=1)

B: Listing of all possible combinations of phosphopeptides using m and n values

For the precursor ion with m/z = 1046.52, it contains two possible phosphorylated amino acids at position 4 and 8 of the sequence (i.e., n = 2), and one phosphate group (i.e., m = 1). Hence, two possible phosphopeptides can be deduced according to C_1^2.

For precursor ion: m/z = 798.31, z = 2
The possible peptides at m/z = 798.31 are:

Peptide	protein3_[1]KDSDDEEEVVHVD[13]	
1	protein3_[1]KDSpDDEEEVVHVD[13]	(z=2)

For precursor ion: m/z = 861.6, z = 3
The possible peptides at m/z = 861.6 are:

Peptide	protein2_[1]LCDFGVSGQLIDSMANSFVGTR[22]	
1	protein2_[1]LCDFGVSpGQLIDSpMANSFVGTR[22]	(z=3)
2	protein2_[1]LCDFGVSpGQLIDSMANSpFVGTR[22]	(z=3)
3	protein2_[1]LCDFGVSpGQLIDSMANSFVGTpR[22]	(z=3)
4	protein2_[1]LCDFGVSGQLIDSpMANSpFVGTR[22]	(z=3)
5	protein2_[1]LCDFGVSGQLIDSpMANSFVGTpR[22]	(z=3)
6	protein2_[1]LCDFGVSGQLIDSMANSpFVGTpR[22]	(z=3)

For precursor ion: m/z = 1046.52, z = 2
The possible peptides at m/z = 1046.52 are:

Peptide	protein4_[1]DNRSQVETEDLILKPGVV[18]	
1	protein4_[1]DNRSpQVETEDLILKPGVV[18]	(z=2)
2	protein4_[1]DNRSQVETpEDLILKPGVV[18]	(z=2)

For precursor ion: m/z = 1101.48, z = 2
The possible peptides at m/z = 1101.48 are:

Peptide	protein5_[1]EKKEFLEPDSWETLDQQ[17]	
1	protein5_[1]EKKEFLEPDSpWETLDQQ[17]	(z=2)
2	protein5_[1]EKKEFLEPDSWETpLDQQ[17]	(z=2)

C: Listing of the possible phosphopeptides using mass information of product ion series

In the following, the first row contains the sequence of the most probable phosphopeptide, resulting in the identified phosphoprotein. The peak ratio value shown in the second row implies the confidence of the identified phosphoprotein. Within the allowable mass tolerance (i.e., 1 Da), a peak ratio value of 1 indicates a perfect match of the mass data of product ions series between experimental (see data in Column 1 of the table below) and deduced (Column 2 of the same table). Column 3 shows the ion type (for example, b[1--3] means b3) for each respective product ion, Column 4 is the predicted charge state, and the predicted peptide sequence of each product ion is listed in the last column of the table.

For precursor ion: m/z = 798.31, z = 2

1 The possible peptide is: protein3_[1]KDSpDDEEEVVHVD[13] (z=2)
The peak ratio is 1.000

The identified sequences are:

411.15 -->	411.16142 -->	b[1--3] [z=1]	KDSp
526.17 -->	526.18791 -->	b[1--4] [z=1]	KDSpD
526.17 -->	526.54956 -->	b[1--13] [z=3]	KDSpDDEEEVVHVD
641.20 -->	641.21440 -->	b[1--5] [z=1]	KDSpDD
770.24 -->	770.25699 -->	b[1--6] [z=1]	KDSpDDE
899.29 -->	899.29958 -->	b[1--7] [z=1]	KDSpDDEE
1028.33 -->	1028.34217 -->	b[1--8] [z=1]	KDSpDDEEE
1127.40 -->	1127.41058 -->	b[1--9] [z=1]	KDSpDDEEEV
1226.47 -->	1226.47899 -->	b[1--10] [z=1]	KDSpDDEEEVV
1363.52 -->	1363.53790 -->	b[1--11] [z=1]	KDSpDDEEEVVH
682.76 -->	682.27292 -->	b[1--11] [z=2]	KDSpDDEEEVVH
732.29 -->	731.80713 -->	b[1--12] [z=2]	KDSpDDEEEVVHV
370.15 -->	370.17703 -->	y[1--3] [z=1]	DVH
469.22 -->	469.24544 -->	y[1--4] [z=1]	DVHV
568.29 -->	568.31385 -->	y[1--5] [z=1]	DVHVV
697.33 -->	697.35644 -->	y[1--6] [z=1]	DVHVVE
826.37 -->	826.39903 -->	y[1--7] [z=1]	DVHVVEE
955.41 -->	955.44162 -->	y[1--8] [z=1]	DVHVVEEE

1070.44 --> 1070.46811 --> y[1--9] [z=1] DVHVVEEED
1185.47 --> 1185.49460 --> y[1--10] [z=1] DVHVVEEEDD
1352.50 --> 1352.52663 --> y[1--11] [z=1] DVHVVEEEDDSp
///

For precursor ion: m/z = 861.6, z = 3

1 The possible peptide is: protein2_[1]LCDFGVSGQLIDSpMANSpFVGTR[22] (z=3)
The peak ratio is 0.708

The identified sequences are:

740.89 --> 740.55677 --> b[1--6] [z=1] LCDFGV
1013.15 --> 1012.66884 --> b[1--9] [z=1] LCDFGVSGQ
1126.31 --> 1125.75290 --> b[1--10] [z=1] LCDFGVSGQL
1126.31 --> 1125.59640 --> b[1--19] [z=2] LCDFGVSGQLIDSpMANSpFV
1239.46 --> 1238.83696 --> b[1--11] [z=1] LCDFGVSGQLI
507.08 --> 506.83839 --> b[1--9] [z=2] LCDFGVSGQ
507.08 --> 507.63712 --> b[1--13] [z=3] LCDFGVSGQLIDSp
563.66 --> 563.38042 --> b[1--10] [z=2] LCDFGVSGQL
620.23 --> 619.92245 --> b[1--11] [z=2] LCDFGVSGQLI
1345.23 --> 1344.54137 --> y[1--11] [z=1] RTGVFSpNAMSpD
1230.14 --> 1229.51488 --> y[1--10] [z=1] RTGVFSpNAMSp
1149.24 --> 1149.51488 --> y[1--10] [z=1] RTGVFSpNAMS
1063.08 --> 1062.48285 --> y[1--9] [z=1] RTGVFSpNAM
931.89 --> 931.44236 --> y[1--8] [z=1] RTGVFSpNA
860.81 --> 860.40525 --> y[1--7] [z=1] RTGVFSpN
860.81 --> 861.46209 --> y[1--22] [z=3] RTGVFSpNAMSpDILQGSVGFDCL
746.81 --> 746.36232 --> y[1--6] [z=1] RTGVFSp
579.66 --> 579.33029 --> y[1--5] [z=1] RTGVF
432.48 --> 432.26188 --> y[1--4] [z=1] RTGV
648.72 --> 647.96862 --> y[1--17] [z=3] RTGVFSpNAMSpDILQGSV
///

2 The possible peptide is: protein2_[1]LCDFGVSGQLIDSpMANSFVGTpR[22] (z=3)
The peak ratio is 0.625

The identified sequences are:

740.89 --> 740.55677 --> b[1--6] [z=1] LCDFGV
1013.15 --> 1012.66884 --> b[1--9] [z=1] LCDFGVSGQ
1126.31 --> 1125.75290 --> b[1--10] [z=1] LCDFGVSGQL
1239.46 --> 1238.83696 --> b[1--11] [z=1] LCDFGVSGQLI
507.08 --> 506.83839 --> b[1--9] [z=2] LCDFGVSGQ
507.08 --> 507.63712 --> b[1--13] [z=3] LCDFGVSGQLIDSp
563.66 --> 563.38042 --> b[1--10] [z=2] LCDFGVSGQL
620.23 --> 619.92245 --> b[1--11] [z=2] LCDFGVSGQLI
1345.23 --> 1344.54137 --> y[1--11] [z=1] RTpGVFSNAMSpD
1230.14 --> 1229.51488 --> y[1--10] [z=1] RTpGVFSNAMSp
1149.24 --> 1149.51488 --> y[1--10] [z=1] RTpGVFSNAMS
1063.08 --> 1062.48285 --> y[1--9] [z=1] RTpGVFSNAM

931.89 --> 931.44236 --> y[1--8] [z=1] RTpGVFSNA
860.81 --> 860.40525 --> y[1--7] [z=1] RTpGVFSN
860.81 --> 861.46209 --> y[1--22] [z=3] RTpGVFSNAMSpDILQGSVGFDCL
746.81 --> 746.36232 --> y[1--6] [z=1] RTpGVFS
648.72 --> 647.96862 --> y[1--17] [z=3] RTpGVFSNAMSpDILQGSV
///
3 The possible peptide is: protein2_[1]LCDFGVSGQLIDSMANSpFVGTpR[22] (z=3)
The peak ratio is 0.500

The identified sequences are:

740.89 --> 740.55677 --> b[1--6] [z=1] LCDFGV
1013.15 --> 1012.66884 --> b[1--9] [z=1] LCDFGVSGQ
1126.31 --> 1125.75290 --> b[1--10] [z=1] LCDFGVSGQL
1239.46 --> 1238.83696 --> b[1--11] [z=1] LCDFGVSGQLI
507.08 --> 506.83839 --> b[1--9] [z=2] LCDFGVSGQ
563.66 --> 563.38042 --> b[1--10] [z=2] LCDFGVSGQL
620.23 --> 619.92245 --> b[1--11] [z=2] LCDFGVSGQLI
1345.23 --> 1344.54137 --> y[1--11] [z=1] RTpGVFSpNAMSD
1230.14 --> 1229.51488 --> y[1--10] [z=1] RTpGVFSpNAMS
860.81-->861.46209-->y[1--22][z=3] RTpGVFSpNAMSDILQGSVGFDCL
746.81 --> 746.36232 --> y[1--6] [z=1] RTpGVFS
648.72 --> 647.96862 --> y[1--17] [z=3] RTpGVFSpNAMSDILQGSV
507.08 --> 506.22515 --> y[1--8] [z=2] RTpGVFSpNA
///
4 The possible peptide is: protein2_[1]LCDFGVSpGQLIDSMANSpFVGTR[22] (z=3)
The peak ratio is 0.458

The identified sequences are:

740.89 --> 740.55677 --> b[1--6] [z=1] LCDFGV
1126.31 --> 1125.59640 --> b[1--19] [z=2] LCDFGVSpGQLIDSMANSpFV
507.08 --> 507.63712 --> b[1--13] [z=3] LCDFGVSpGQLIDS
1149.24 --> 1149.51488 --> y[1--10] [z=1] RTGVFSpNAMS
1063.08 --> 1062.48285 --> y[1--9] [z=1] RTGVFSpNAM
931.89 --> 931.44236 --> y[1--8] [z=1] RTGVFSpNA
860.81 --> 860.40525 --> y[1--7] [z=1] RTGVFSpN
860.81-->861.46209-->y[1--22][z=3] RTGVFSpNAMSDILQGSpVGFDCL
746.81 --> 746.36232 --> y[1--6] [z=1] RTGVFSp
746.81 --> 745.85871 --> y[1--13] [z=2] RTGVFSpNAMSDIL
579.66 --> 579.33029 --> y[1--5] [z=1] RTGVF
432.48 --> 432.26188 --> y[1--4] [z=1] RTGV
648.72 --> 647.96862 --> y[1--17] [z=3] RTGVFSpNAMSDILQGSpV
///
5 The possible peptide is: protein2_[1]LCDFGVSpGQLIDSpMANSFVGTR[22] (z=3)
The peak ratio is 0.375

The identified sequences are:

740.89 --> 740.55677 --> b[1--6] [z=1] LCDFGV
1126.31 --> 1125.59640 --> b[1--19] [z=2] LCDFGVSpGQLIDSpMANSFV

507.08 --> 507.63712 --> b[1--13] [z=3] LCDFGVSpGQLIDS
1149.24 --> 1149.51488 --> y[1--10] [z=1] RTGVFSNAMSp
860.81-->861.46209-->y[1--22][z=3] RTGVFSNAMSpDILQGSpVGFDCL
746.81 --> 745.85871 --> y[1--13] [z=2] RTGVFSNAMSpDIL
579.66 --> 579.33029 --> y[1--5] [z=1] RTGVF
432.48 --> 432.26188 --> y[1--4] [z=1] RTGV
648.72 --> 647.96862 --> y[1--17] [z=3] RTGVFSNAMSpDILQGSpV
///

For precursor ion: m/z = 1046.52, z = 2

1 The possible peptide is: protein4_[1]DNRSpQVETEDLILKPGVV[18] (z=2)
The peak ratio is 1.000

The identified sequences are:
386.17 --> 386.17847 --> b[1--3] [z=1] DNR
681.26 --> 681.26908 --> b[1--5] [z=1] DNRSpQ
780.32 --> 780.33749 --> b[1--6] [z=1] DNRSpQV
909.37 --> 909.38008 --> b[1--7] [z=1] DNRSpQVE
909.37 --> 909.95234 --> b[1--15] [z=2] DNRSpQVETEDLILKP
1139.46 --> 1139.47035 --> b[1--9] [z=1] DNRSpQVETE
1254.48 --> 1254.49684 --> b[1--10] [z=1] DNRSpQVETED
1480.65 --> 1480.66496 --> b[1--12] [z=1] DNRSpQVETEDLI
1593.74 --> 1593.74902 --> b[1--13] [z=1] DNRSpQVETEDLIL
1721.83 --> 1721.84398 --> b[1--14] [z=1] DNRSpQVETEDLILK
371.21 --> 371.23426 --> y[1--4] [z=1] VVGP
499.30 --> 499.32922 --> y[1--5] [z=1] VVGPK
612.39 --> 612.41328 --> y[1--6] [z=1] VVGPKL
725.47 --> 725.49734 --> y[1--7] [z=1] VVGPKLI
838.55 --> 838.58140 --> y[1--8] [z=1] VVGPKLIL
953.58 --> 953.60789 --> y[1--9] [z=1] VVGPKLILD
1082.62 --> 1082.65048 --> y[1--10] [z=1] VVGPKLILDE
1183.67 --> 1183.69816 --> y[1--11] [z=1] VVGPKLILDET
1312.72 --> 1312.74075 --> y[1--12] [z=1] VVGPKLILDETE
1411.78 --> 1411.80916 --> y[1--13] [z=1] VVGPKLILDETEV
1539.84 --> 1539.86774 --> y[1--14] [z=1] VVGPKLILDETEVQ
///

2 The possible peptide is: protein4_[1]DNRSQVETpEDLILKPGVV[18] (z=2)
The peak ratio is 0.750

The identified sequences are:

386.17 --> 386.17847 --> b[1--3] [z=1] DNR
909.37 --> 909.95234 --> b[1--15] [z=2] DNRSQVETpEDLILKP
1139.46 --> 1139.47035 --> b[1--9] [z=1] DNRSQVETpE
1254.48 --> 1254.49684 --> b[1--10] [z=1] DNRSQVETpED
1480.65 --> 1480.66496 --> b[1--12] [z=1] DNRSQVETpEDLI
1593.74 --> 1593.74902 --> b[1--13] [z=1] DNRSQVETpEDLIL
1721.83 --> 1721.84398 --> b[1--14] [z=1] DNRSQVETpEDLILK
371.21 --> 371.23426 --> y[1--4] [z=1] VVGP

499.30 --> 499.32922 --> y[1--5] [z=1] VVGPK
612.39 --> 612.41328 --> y[1--6] [z=1] VVGPKL
725.47 --> 725.49734 --> y[1--7] [z=1] VVGPKLI
838.55 --> 838.58140 --> y[1--8] [z=1] VVGPKLIL
953.58 --> 953.60789 --> y[1--9] [z=1] VVGPKLILD
1082.62 --> 1082.65048 --> y[1--10] [z=1] VVGPKLILDE
1183.67 --> 1183.69816 --> y[1--11] [z=1] VVGPKLILDET
///

For precursor ion: m/z = 1101.48, z = 2

1 The possible peptide is: protein5_[1]EKKEFLEPDSpWETLDQQ[17] (z=2)
The peak ratio is 1.000

The identified sequences are:

604.28 --> 604.95297 --> b[1--14] [z=3] EKKEFLEPDSpWETL
904.48 --> 904.47810 --> b[1--7] [z=1] EKKEFLE
1116.55 --> 1116.55735 --> b[1--9] [z=1] EKKEFLEPD
1469.67 --> 1469.66869 --> b[1--11] [z=1] EKKEFLEPDSpW
1598.71 --> 1598.71128 --> b[1--12] [z=1] EKKEFLEPDSpWE
1812.84 --> 1812.84302 --> b[1--14] [z=1] EKKEFLEPDSpWETL
1927.86 --> 1927.86951 --> b[1--15] [z=1] EKKEFLEPDSpWETLD
390.15 --> 390.16687 --> y[1--3] [z=1] QQD
604.28 --> 604.29861 --> y[1--5] [z=1] QQDLT
1201.46 --> 1201.47903 --> y[1--9] [z=1] QQDLTEWSpD
1298.51 --> 1298.53179 --> y[1--10] [z=1] QQDLTEWSpDP
1427.56 --> 1427.57438 --> y[1--11] [z=1] QQDLTEWSpDPE
1687.71 --> 1687.72685 --> y[1--13] [z=1] QQDLTEWSpDPELF
650.56 --> 649.76986 --> y[1--10] [z=2] QQDLTEWSpDP
///

2 The possible peptide is: protein5_[1]EKKEFLEPDSWETpLDQQ[17] (z=2)
The peak ratio is 0.846

The identified sequences are:

604.28 --> 604.95297 --> b[1--14] [z=3] EKKEFLEPDSWETpL
904.48 --> 904.47810 --> b[1--7] [z=1] EKKEFLE
1116.55 --> 1116.55735 --> b[1--9] [z=1] EKKEFLEPD
1812.84 --> 1812.84302 --> b[1--14] [z=1] EKKEFLEPDSWETpL
1927.86 --> 1927.86951 --> b[1--15] [z=1] EKKEFLEPDSWETpLD
390.15 --> 390.16687 --> y[1--3] [z=1] QQD
604.28 --> 604.29861 --> y[1--5] [z=1] QQDLT
1201.46 --> 1201.47903 --> y[1--9] [z=1] QQDLTpEWSD
1298.51 --> 1298.53179 --> y[1--10] [z=1] QQDLTpEWSDP
1427.56 --> 1427.57438 --> y[1--11] [z=1] QQDLTpEWSDPE
1687.71 --> 1687.72685 --> y[1--13] [z=1] QQDLTpEWSDPELF
650.56 --> 649.76986 --> y[1--10] [z=2] QQDLTpEWSDP
///