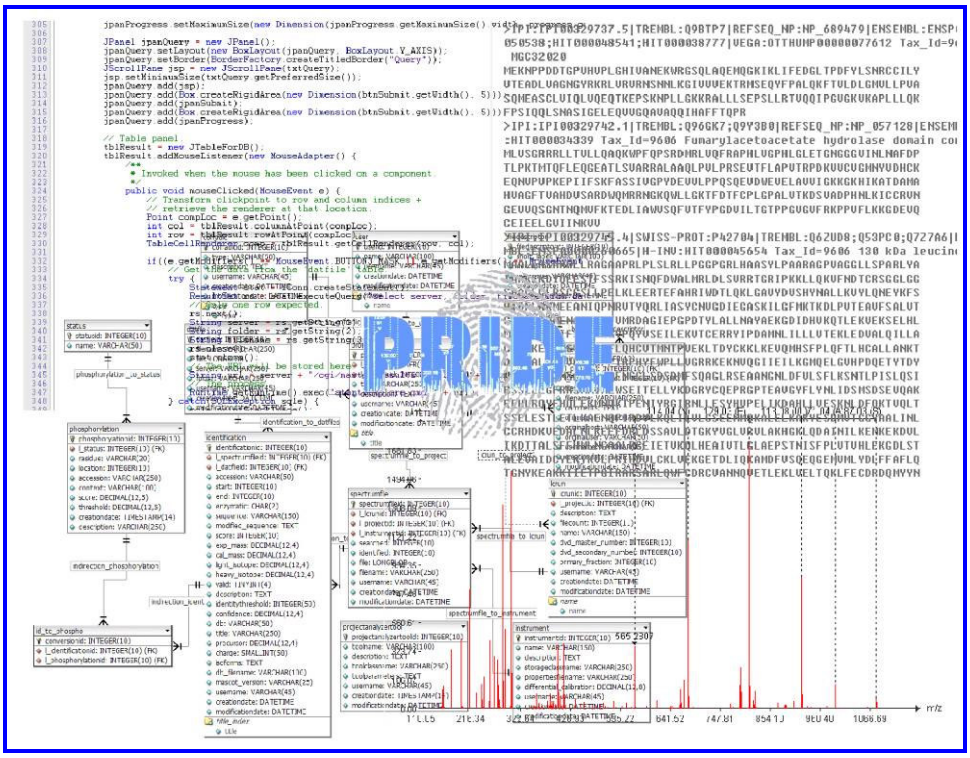
Novel bioinformatics tools assisting targeted peptide-centric proteomics and global proteomics data dissemination.



Lennart MARTENS

Promotor: Prof. Dr. Joël Vandekerckhove
Co-promotor: Prof. Dr. Kris Gevaert

UNIVERSITEIT GENT

VIB
FLANDERS INTERUNIVERSITY INSTITUTE FOR BIOTECHNOLOGY

Novel bioinformatics tools assisting targeted peptide-centric proteomics and global proteomics data dissemination.

Thesis submitted to obtain the degree of Doctor (Ph. D.) in Sciences: Biotechnology

Lennart Martens

2006

Promotor: Prof. Dr. Joël Vandekerckhove

Department of Biochemistry, GE07
Department of Medical Protein Research, VIB09

Co-promotor: Prof. Dr. Kris Gevaert

Department of Biochemistry, GE07
Department of Medical Protein Research, VIB09

“Why is programming fun? What delights may its practitioner expect as his reward?

First is the sheer joy of making things. As the child delights in his mud pie, so the adult enjoys building things, especially things of his own design. I think this delight must be an image of God's delight in making things, a delight shown in the distinctness and newness of each leaf and each snowflake.

Second is the pleasure of making things that are useful to other people. Deep within, we want others to use our work and find it helpful. In this respect, the programming system is not essentially different from the child's first clay pencil holder "for Daddy's office."

Third is the fascination of fashioning complex puzzle-like objects of interlocking moving parts and watching them work in subtle cycles, playing out the consequences of the principles built in from the beginning. The programmed computer has all the fascination of the pinball machine or the jukebox mechanism carried to the ultimate.

Fourth is the joy of always learning, which springs from the nonrepeating nature of the task. In one way or another, the problem is always new, and its solver learns something: sometimes practical, sometimes theoretical, and sometimes both.

Finally there is the delight of working in such a tractable medium. The programmer, like the poet, works only slightly removed from pure thought-stuff. He builds his castles in the air, from air, creating by exertion of the imagination. Few media of creation are so flexible, so easy to polish and rework, so readily capable of realizing grand conceptual structures.

Yet the program construct, unlike the poet's words, is real in the sense that it moves and works, producing visible outputs seperate from the construct itself. It prints results, draws pictures, produces sounds, moves arms. The magic of myth and legend has come true in our time. One types the correct incantation on the keyboard, and a display screen comes to life, showing things that never were nor could be.

Programming then is fun because it gratifies creative longings built deep within us and delights sensibilities we have in common with all men."

- Frederick P. Brooks, ***The Mythical Man-Month***

Acknowledgements

The undertaking of a Ph. D. is not the lonely work it is often believed to be. Rather the opposite, a doctoral dissertation typically relies on the help and encouragement of many. This work then is no exception as I am fortunate enough to be able to thank many individuals without whom this dissertation would not have come to pass, and – perhaps more importantly – without whom it simply would not have been such an enjoyable experience.

First and foremost I would like to extend my thanks to my promotor, Prof. Dr. Joël Vandekerckhove. A brilliant scientist and a very amiable man, Joël always gave me the necessary freedom to pursue my scientific interests and consistently supported me with his invaluable advice.

The input from Prof. Dr. Kris Gevaert, who bore the brunt of the burden of supervising my work, has been instrumental in nearly all aspects of this dissertation. Kris has time and again shown his dedication to his role as a supervisor by his encouragements, his expert advice and by sharing his insights and ideas.

The entire staff of the proteomics lab in Ghent has contributed in no small measure both to the work presented in this thesis as well as to the pleasant and vibrant atmosphere in the lab. Dr. Xavier Hanoulle and Dr. Jef Pinxteren have repeatedly transferred considerable aliquots of their impressive knowledge during our interesting conversations. I also consider it a rare privilege to have been able to work with such enthousiastic and capable graduate students as Petra Van Damme, Bart Ghesquière and Francis Impens. Apart from filling my ms_lims database with the results of their experiments, their bug spotting abilities and feature requests allowed me to constantly refine and extend my software programs. Marc Goethals, José Van Damme, Hans Caster, Hans Demol, An Staes, Evy Timmerman and Koen Hugelier have been pivotal in expanding my understanding of the instruments they so skillfully manage and operate and they have proven themselves to be quick-learning, patient and understanding users of my software.

My stay as a Marie Curie fellow at the European Bioinformatics Institute and the ongoing collaborations that ensued have indebted me firstly to Dr. Rolf Apweiler. It was Rolf who first sparked my interest in the work done at the EBI through his own enthousiasm and energy. Always aware of the needs and achievements of his staff, Rolf to this day continuously impresses me by flawlessly combining a mind-bogglingly busy schedule with perfect management and communication skills.

My thanks need also be extended to Henning Hermjakob, my team leader at the EBI. Henning has proven himself to be a great supervisor, a fine friend and has even granted me the opportunity of staying on his boat while he was away.

PRIDE would not be the stable, production-grade software system it is today without the tireless efforts of Phil Jones, one of the finest and most dedicated programmers I have had the pleasure to meet. I would also like to thank Phil for proofreading this dissertation.

Richard Côté, also developer on the PRIDE team, has proven invaluable by applying his detailed knowledge of ontologies to provide PRIDE with one of the single most useful software components in the field of proteomics today: his ontology lookup service.

The friendship and support of Samuel Kerrien, Michael Müller, Markus Brosch, Daniela Wieser, Lawrence Bower, Robert Petryszak, Dr. John Garavelli and Kai Runte at the EBI have always made my stays there extremely enjoyable and informative.

One of the great things about being a scientist is that you get the opportunity to build collaborations and friendships with bright and engaging people from all over the world. In this category, my thanks go firstly to Kristian Flikka from the University of Bergen in Norway. His energy, expertise and unfailing sense of humor have made the sharing of an office these last few months a real pleasure and enabled the creation of exciting synergies in our reseach.
Prof. Dr. Alexey Nesvizhskii of the University of Michigan at Ann Arbor, whom I first met at the Plasma Proteome Project jamboree in Ann Arbor, was kind enough to provide me with his considerable expertise and knowledge through his assistance in writing the 'raw data' paper, for which I am very thankful.
Dr. Thomas W. Blackwell of the University of Michigan at Ann Arbor deserves my thanks for helping me establish the statistics employed in the DiffAnalysisGUI application in ms_lims and for introducing me to robust statistics in the process.

Marc Portier earns my gratitude because he successfully conveyed his love of open source programming and readily shared his highly incisive intelligence during our conversations.

Life is of course more than work, but for a graduate student the distinction sometimes seems to fade. My parents and brother and my family-in-law repeatedly helped me across this fine line for some relaxing afternoons during weekends and holidays and often showed an interest in my work and progress for which they also receive a very sincere 'thank you' here.

One of the most effective ways for me to shake off the programming dust gathered over many hours of staring at a fluorescent or LCD monitor has been martial arts training. I am therefore greatly indebted to Tino Brebels and Roel Zoons for their guidance and support while providing me with the stress-relieving and character-building training sessions that I have come to value as an important part of my life.

It is important to note that most people will only remember the first and last persons mentioned in the acknowledgements. It is for this reason only that I mention here at the very end my wife Leen Kestens, who has stood by me through all these years, displaying endless and almost inhuman patience and understanding. For consistently being the light of my life and the dearest friend imaginable, I would like to conclude by extending my most cordial thanks to her.

Table of contents

List of publications

Publications represented in this work *(in chronological order)*

Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR, Vandekerckhove J, 'Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides', **2003**, *Nature Biotechnology* **21**:566-569 (IP: 11.310).

Gevaert K, Ghesquière B, Staes A, Martens L, Van Damme J, Thomas GR, Vandekerckhove J, 'Reversible labeling of cysteine-containing peptides allows their chromatographic isolation for non-gel proteome studies', **2004**, *Proteomics* **4**:897-908 (IP: 5.766).

Martens L, Van Damme P, Van Damme J, Staes A, Timmerman E, Ghesquière B, Thomas GR, Vandekerckhove J, Gevaert K, 'The human platelet proteome mapped by peptide-centric proteomics: a functional protein profile', **2005**, *Proteomics* **5**:3193-3204 (IP: 5.766).

Martens L, Nesvizhskii AI, Hermjakob H, Adamski M, Omenn GS, Vandekerckhove J, Gevaert K, 'Do we want our data raw? Including binary mass-spectrometry data in public proteomics data repositories', **2005**, *Proteomics* **5**:3501-3505 (IP: 5.766).

Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R, 'PRIDE: The PRoteomics IDEntifications database', **2005**, *Proteomics* **5**:3537-3545 (IP: 5.766).

Gevaert K, Staes A, Van Damme J, De Groot S, Hugelier K, Demol H, Martens L, Goethals M, Vandekerckhove J, 'Global phosphoproteome analysis on human HepG2 hepatocytes using reversed-phase diagonal LC', **2005**, *Proteomics* **5**:3589-3599 (IP: 5.766).

Martens L, Vandekerckhove J, Gevaert K, 'DBToolkit: processing sequence databases for enhanced peptide identification in peptide-centric proteome analyses', **2005**, *Bioinformatics* **21**:3584-3585 (IP: 5.742).

Van Damme P, Martens L, Van Damme J, Hugelier K, Staes A, Vandekerckhove J, Gevaert K, 'Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis', **2005**, *Nature Methods* **2**:771-777 (IP: to be determined).

Jones P, Côté RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R, 'PRIDE: A Public Repository of Protein and Peptide Identifications for the Proteomics Community', **2006**, *Nucleic Acids Research* **34** (database issue):D659-D663 (IP: 7.260).

Martens L, Flikka K, Gevaert K, Vandekerckhove J, Eidhammer I, 'Improving the Reliability and Throughput of Mass Spectrometry Based Proteomics by Spectrum Quality Filtering', **2006**, *Proteomics* **6**:2086-2094.

Other publications *(in chronological order)*

Gevaert K, Demol H, Martens L, Hoorelbeke B, Puype M, Goethals M, Van Damme J, De Boeck S, Vandekerckhove J, 'Protein identification based on matrix-assisted laser desorption/ionization-post-source decay-mass spectrometry', **2001**, *Electrophoresis* **22**:1645-1651 (IP: 4.282).

Gevaert K, Van Damme J, Goethals M, Thomas GR, Hoorelbeke B, Demol H, Martens L, Puype M, Staes A, Vandekerckhove J, 'Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis – Identification of more than 800 Escherichia coli proteins', **2002**, *Molecular and Cellular Proteomics* **1**:896-903 (IP: 8.316).

Staes A, Demol H, Van Damme J, Martens L, Vandekerckhove J, Gevaert K, 'Global differential non-gel proteomics by quantitative and stable labeling of tryptic peptides with oxygen-18', **2004**, *Journal of Proteome Research* **3**:786-791 (IP: 6.917).

Gevaert K, Van Damme P, Martens L, Vandekerckhove J, 'Diagonal reverse-phase chromatography applications in peptide-centric proteomics; ahead of catalogue-omics?', **2005**, *Analytical Biochemistry* **345**:18-29 (IP: 2.370).

Adamski M, Blackwell T, Menon R, Martens L, Hermjakob H, Taylor C, Omenn GS and States DJ, 'Data Management in the Pilot Phase of the HUPO Plasma Proteome Project', **2005**, *Proteomics* **5**:3246-3261 (IP: 5.766).

Stephan C, Hamacher M, Bluggel M, Korting G, Chamrad D, Scheer C, Marcus K, Reidegeld KA, Lohaus C, Schafer H, Martens L, Jones P, Muller M, Auyeung K, Taylor C, Binz PA, Thiele H, Parkinson D, Meyer HE, Apweiler R., '5(th) HUPO BPP Bioinformatics Meeting at the European Bioinformatics Institute in Hinxton, UK - Setting the Analysis Frame', **2005**, *Proteomics* **5**:3560-3562 (IP: 5.766).

Reidegeld KA, Hamacher M, Meyer HE, Stephan C, Blüggel M, Körting G, Chamrad D, Scheer C, Thiele H, Taylor C, Müller M, Apweiler R, Jones P, Martens L, 'The HUPO Brain Proteome Project', **2006**, *European Pharmaceutical Review* **1**:33-38.

Hamacher M, Stephan C, Bluggel M, Chamrad D, Korting G, Martens L, Muller M, Hermjakob H, Parkinson D, Dowsey A, Reidegeld KA, Marcus K, Dunn MJ, Meyer HE, Apweiler R, 'The HUPO Brain Proteome Project Jamboree: Centralised summary of the pilot studies', **2006**, *Proteomics* **6**:1719-1721 (IP: 5.766).

Gevaert K, Pinxteren J, Demol H, Hugelier K, Staes A, Van Damme J, Martens L, Vandekerckhove J, 'Four Stage Liquid Chromatographic Selection of Methionyl Peptides for Peptide-Centric Proteome Analysis: The Proteome of Human Multipotent Adult Progenitor Cells', *Journal of Proteome Research*, *in press* (IP: 6.917).

Abbreviations

1:1	one-to-one mapping
1:n	one-to-many mapping
1D	one-dimensional
2D	two-dimensional
2D-PAGE	two-dimensional polyacrylamide gel electrophoresis
AA	amino acid
API	application programmers interface
Arg-C	endoproteinase Arg-C
BLOB	binary large object
CID	collision-induced dissocation
COFRADIC	combined fractional diagonal chromatography
C-terminal	carboxyterminal
CVS	concurrent versioning system
dd.	de dato
DTNB	5,5-dithiobis(2-nitrobenzoic acid)
ESI	electrospray ionisation
EST	expressed sequence tags
false (+)	false positive
GNU	GNU is not Unix
GPL	General Public License
GUI	graphical user interface
GZIP	GNU zip
HPLC	high-performance liquid chromatography
HPPP	HUPO plasma proteome project
HTML	hypertext mark-up language
HUGO	human genome organisation
HUPO	Human Proteome Organisation
I/O	input-output
ICAT	isotope coded affinity tag
IEF	isoelectric focusing
JDBC	Java database connectivity
JDK	Java development kit
JRE	Java runtime environment
LC	liquid chromatography

LIMS	laboratory information management system
m/z	mass-to-charge ratio
MALDI	matrix-assisted laser desorption and ionisation
MIAPE	minimal information about a proteomics experiment
MS	mass spectrometry
MS/MS	tandem MS
MudPIT	multidimensional proteome identification technique
n:m	many-to-many mapping
N-terminal	aminoterminal
ORF	open reading frame
PCR	polymerase chain reaction
pI	isoelectric point
PROBE	proteomics unit of the University of Bergen
PSD	post-source decay
PSI	Proteomics Standards Initiative
RDBMS	relational database management system
RP	reverse-phase
SDS	sodium dodecyl sulphate
SQL	structured query language
TCEP	tris(2-carboxyethyl)phosphine
TFA	tri-fluoro acetate
TNB	Thionitrobenzoyl
TNBS	2,4,6-trinitrobenzenesulfonic acid
TOF	time-of-flight
UML	unified modeling language
UV	Ultraviolet

Foreword

The field of proteomics is quickly maturing into what might very well be its golden age. Indeed, over the past several years, the foundations of this promising branch of the life sciences have been firmly established. The first tier was laid down by three technological improvements: the invention of the two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) method for separating proteins from a complex mixture, the development of automated search algorithms that could identify a protein from a sequence database by the (fragment) masses of its peptide ions, and the improvement of mass spectrometry instrumentation for the analysis of large biomolecules.
The second tier built primarily upon the completion of the human genome sequencing projects. The resulting wealth of genomic sequence information was combined with expressed sequence tag (EST) data that had been gathered in parallel to help prediction algorithms in their search for open reading frames (ORF's). These massive efforts in turn fed their data to protein sequence databases, greatly increasing the amount of primary protein structure information available to the abovementioned identification algorithms. At this point in time however, proteomics remained a métier rather than a protocol step, with the outcome of an experiment more dependent on the hands and minds that executed it than on the instrumentation and methodology applied. Correspondingly, the output of the average proteomics lab was typically limited to several identified proteins per week. As the instrumentation became ever more sophisticated however, the community started to develop novel ways to tackle old problems and over a relatively short period of time several groups independently published exciting new techniques that went squarely beyond the protein and focused on the constituent peptides instead. This third tier of novel peptide-centric techniques carried enormous potential, being readily automatable, highly sensitive and often very adaptable to suit specific research needs. Indeed, the new techniques combined with the dramatically improved instrumentation in a truly synergistic way and together they metamorphosed the field almost overnight. Where data production had previously resembled a rather sluggish trickling of identifications, data now streamed from the instruments in veritable torrents.
In order to maximize the information contained in this data flood, novel bioinformatics applications were desperately needed as both existing data management software as well as available data processing software were lagging behind these new developments.
On a global level, dissemination and validation of the identifications obtained posed another problem: whereas the identifications obtained could previously be published in neat, well-annotated tables inserted in the main text of a paper, the several hundreds of protein identifications routinely obtained using the new techniques were necessarily exiled to the supplementary information, losing much of their informative aspect in the process.

In this work some contributions to alleviate these software requirements will be discussed. These applications cover solutions to a wide range of bottlenecks, from the processing of sequence databases (section 2.1), the management and analysis of the information flow (section 2.2), spectrum quality filtering (section 2.3) and global proteomics data dissemination (section 2.4). Finally, a summary of all of the above tools is given through an applied example to catalogue the most complete human platelet proteome to date in section 2.5.

1. Introduction

1.1. From genomics to proteomics

1.1.1. The genomic era

The independent completion of the sequencing of the human genome by both Celera Corporation [Venter 2001] and the Human Genome Organisation (HUGO) [Lander 2001] has been one of the milestone achievements of biology. The corresponding availability of automated, high-throughput sequencers allowed the completion of several genomes of model organisms since then [Aparicio 2002, Waterston 2002, Gibbs 2004, Hillier 2004, Mikkelsen 2005].
In turn, the availability of whole-genome sequences started off large-scale searches for open reading frames (ORF's). Both *in silico* efforts and mRNA sequencing contributed greatly to this end and today we have a very good estimate of the total number of expected genes in the human genome [International Human Genome Sequencing Consortium 2004].
One thing that becomes clear when one looks at the relative scarcity of coding sequences in the genome is that a study of the genome by itself cannot account for an explanation of cellular functioning [Claverie 2001, Harrison 2002]. Thus, although a genome provides the all-important basis for a better understanding of a living organism, it cannot by itself provide an explanation for the actual diversity and adaptability evident in all life. Rather, one has to look into the RNA and protein content of cells in order to get an idea of how the cell works at any given time.

1.1.2. Discovering the proteome

The term 'proteome' is generally attributed to Mark Wilkins, who coined the term at the Siena Conference in 1994 [Wasinger 1995]. Through borrowing the semantics from the 'genome' term, it becomes clear that the scope of research is very similar in both cases, yet the topics differ. Indeed, in order to understand a living cell at the molecular level it is imperative to analyze its protein content. Analyzing the proteome presents a more daunting challenge than analyzing the genome: apart from spanning an extremely large concentration range (at least 10 orders of magnitude in plasma [States 2006]) it is both highly dynamic in concentration as well as in modification state. Indeed, even though cells share the same genome, their proteomes can differ markedly [Collins 2001]. Additionally, protein sequences are not easily duplicated to large copy-numbers as is the case for nucleic acid sequences through the application of the polymerase chain reaction (PCR).
The most popular technique today for studying the proteome is by mass spectrometry, which relies on separating charged ions by their mass-to-charge ratio (m/z).

1.1.3. Mass spectrometry

In general, a mass spectrometer can be broken down into three parts: an ion source, an m/z analyzer and a detector. The latter is usually a specific type of electron multiplier.

Due to the high amplification that is typically required most modern instruments use a type of microchannel plate detector.
Additional refinements for peptide/protein sequence analysis include so-called tandem-MS or MS/MS instruments which are capable of more than one round of mass spectrometry. In this technique one mass spectrometer isolates a peptide of a particular m/z while a second mass spectrometer is used to catalog fragment ions resulting after induced or spontaneous fragmentation.
Different sources, m/z analyzers and fragmentation mechanisms are outlined below.

1.1.3.1. Ion sources

When applied to biomolecular compounds, the ion source of the mass spectrometer can typically take two forms: a Matrix-Assisted Laser Desorption and Ionisation (MALDI) source [Karas 1988, Tanaka 1988] or an ElectroSpray Ionisation (ESI) source [Fenn 1989].
In a MALDI source, energy from laser light is converted into kinetic energy of the irradiated molecules/ions. This light is directed towards a metallic target plate on which the analyte has been crystallized in the presence of so-called matrix molecules. The laser light itself is typically derived from a N_2 laser generating UV light with a wavelength of 337 nm. Some often-used matrix compounds are α-cyano-4-hydroxycinnamic acid for peptide analytes [Beavis 1990] and sinapinic acid (3,5-dimethoxy-4-hydroxycinnamic acid) for proteins [Beavis, 1989]. Crystallization is usually performed in highly organic solvents and in the presence of 0.1% tri-fluoroacetate (TFA).
The actual mechanisms leading to desorption and ionisation are subject of debate [Karas 1996, Zhao 1997,Jørgensen 1998, Wong 1998, Zenobi 1998], yet it is thought to rely on efficient absorption of the laser energy by the matrix molecules, which ultimately convert it into kinetic energy. This theory explains why a high molar excess of matrix molecules is required to obtain efficient desorption of the analyte. Ionisation might, according to one hypothesis, occur in the gas phase by proton transfer between the acidic matrix ions and the basic residues of the analyte (lysine, arginine or histidine).
The principle of a MALDI source is depicted according to this theory in figure 1.

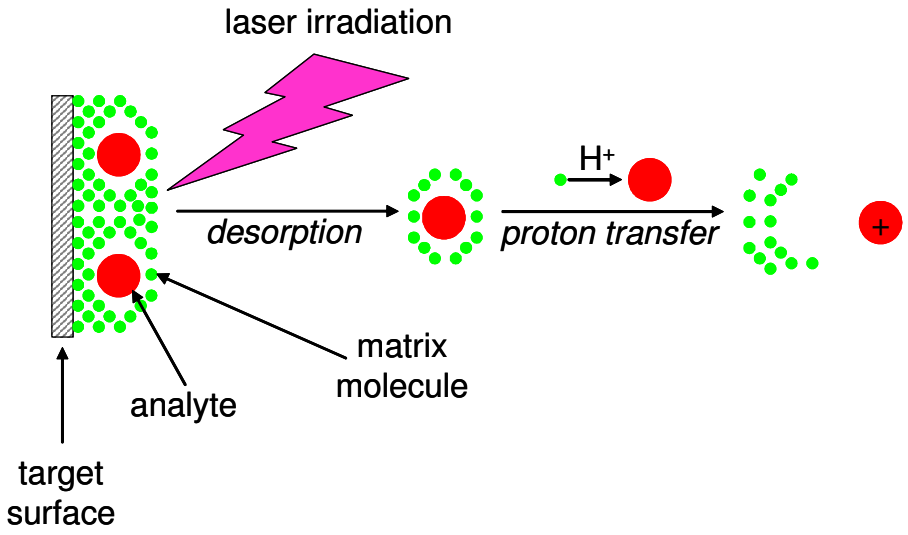


Figure 1: General principle of MALDI.

Electrospray ionisation follows a completely different approach to generating gas-phase ions from an analyte. First and foremost, an ESI source starts from the analyte in solution, requiring a continuous flow to be provided at the inlet. Second, the phase transition here is evaporation under atmospheric pressure rather than the sublimation in a vacuum used for a MALDI source.
The solution in which the analyte is dissolved is composed of compounds that are more volatile than the analyte itself and is typically acidified for peptide analysis, causing the peptides to become protonated. By passing this fluid through a slightly warm (50-60 °C) conductive needle over which a voltage of 3 to 5 kV is applied, the positive charges of the peptides will be pushed out of the needle tip, forming a Taylor cone [Taylor 1964] in the process. The resulting aerosol of tiny, charged droplets (the actual *electrospray*) continues to evaporate (sometimes assisted by blowing a neutral carrier gas along the conductive needle and through the source) and migrates towards the counter electrode. As the droplets evaporate, the ever increasing charge density destabilizes the droplets which fall apart in ever smaller droplets (i.e. the Coulomb explosion phenomenon), ultimately creating multiply charged, individual gas-phase analyte ions. The counter electrode and mass analyzer inlet are often positioned at right angles to the tip of the needle to filter out any uncharged ions (which, blind to the electrical field, continue in a straight path).
The principle of an ESI source is depicted in figure 2.

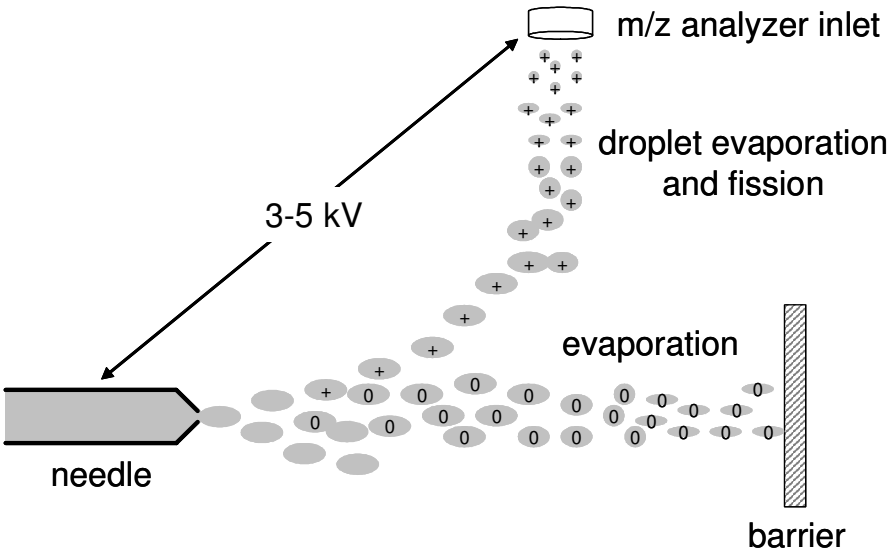


Figure 2: General principle of an ESI source.

It is important to highlight that, under proper conditions, MALDI samples can be archived for later re-use whereas ESI samples are spent over the analysis period. Additionally, both sources ultimately perform an incomplete sampling of the analyte molecules present as MALDI ionization is a competitive process in which the available protons are preferentially captured by some analytes. This *ionisation-quenching* phenomenon can completely obscure an instrument fed by a MALDI-source to certain peptides present in the original matrix crystals. For ESI sources, the limited amount of time during which any particular ion elutes presents the greatest hurdle towards full analyte coverage. Most peptide sequences will be ionized in the source, but the mass spectrometer can only provide a detailed (fragmentation) analysis of one of multiple ions that are presented at the inlet simultaneously[1].

1.1.3.2. Mass/charge analyzers

Analyzers measure the mass-to-charge (m/z) ratios of the ions presented to them. There are two kinds of mass analyzers: those that employ electrical fields and those that rely on magnetic fields. The former group is composed of time-of-flight (TOF), quadrupole and ion trap analyzers while the latter group comprises magnetic sector and ion cyclotron analyzers.

Today, the four analyzers in common use on mass spectrometers are the time-of-flight, quadrupole, ion trap and Fourier transform ion cyclotron resonance analyzers[2].

The analyzers used in the work discussed here will be outlined next.

A TOF analyzer is essentially a long tube with the inlet from the ion source at one end and a detector at the other. The ions generated in the source are briefly (appr. 1 cm)

[1] Certain modern machines allow the selection and fragmentation of multiple precursors simultaneously thus partially overcoming this problem.

[2] Very recently, a fifth type of analyzer was introduced by Thermo Finnigan: the orbitrap.

exposed to an electrical field (typically 15-30 kV), which transfers kinetic energy to the ions according to the following formula:

$$E_k = q \cdot V = \frac{mv^2}{2}$$

With q the charge of the ion, V the voltage applied, m the mass of the ion and v the speed of the ion. The ions are then allowed to fly through a field-free tube of known length. Applying the equation from Newton's dynamics, we can resolve m and q for time (t) and distance (x):

$$x = v \cdot t$$

$$\frac{m}{q} = \frac{2 \cdot V \cdot t^2}{x^2}$$

Since the length of the TOF tube and the voltage applied are known and the actual flight time itself can be determined, it is possible to resolve mass over charge (m/q, or m/z as it is more commonly written).
Ions of the same mass often show a distribution of kinetic energies in a linear TOF analyser after being accelerated, however. This is attributed to three possible causes [Amft 1997] and can be countered through the use of a so-called reflectron field and a delayed (discontinuous) extraction of ions from the ion source.
The principle of a reflectron TOF mass spectrometer builds on the basic layout of a linear instrument, but adds a reflecting electrical field (reflectron) at the end of the flight tube. This reflectron operates at a higher voltage than the extractor field [Mamyrin 1973] and thus reflects ion trajectories back in such a way that ions with equal m/z but different energies are re-aligned and reach the detector simultaneously [Cotter 1997].
A reflectron TOF analyzer is depicted in figure 3.

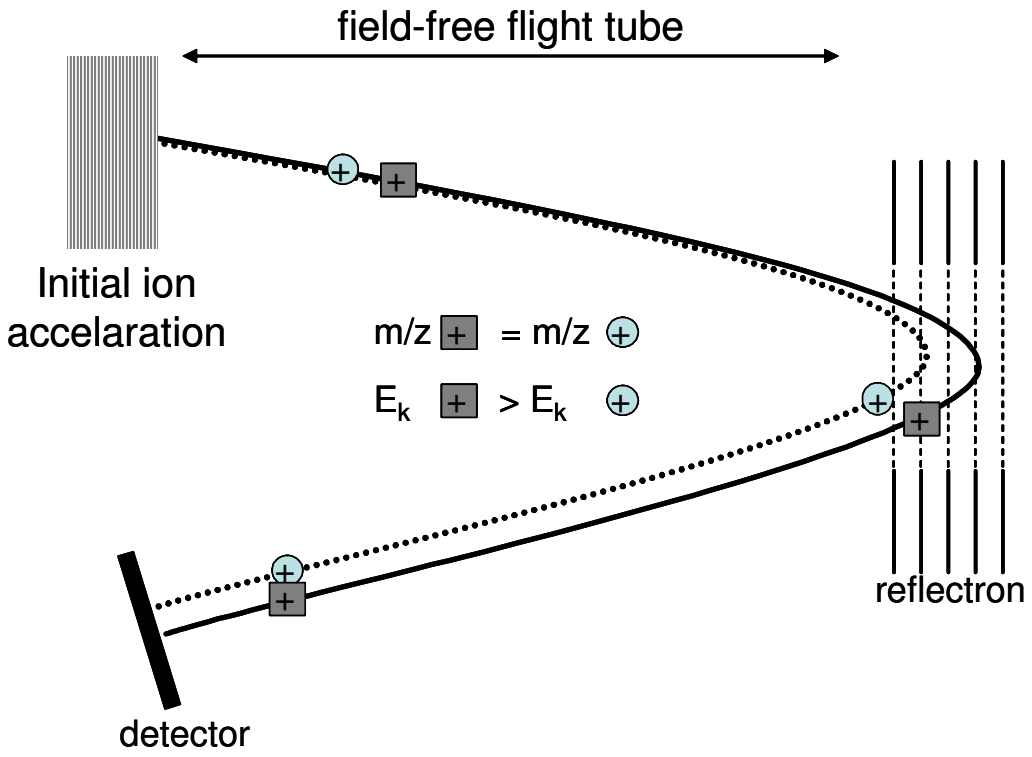


Figure 3: Correction on the kinetic energy distributions by a reflectron field.

A second type of mass analyzer is the quadrupole [Paul 1953]. It analyzes ions by applying an oscillating electrical field on four parallel conductors (the actual quadrupole), which have been connected two-by-two. The electrical field itself is generated by the application of a direct-current (U) and a radiofrequency alternating current ($V.\cos(\omega.t)$). Ions brought into this field will oscillate in the two dimensions orthogonal to the direction of movement. The oscillation itself is dependent on the mass and charge of the ion and will only yield a finite amplitude for certain m/z values given V, U and ω. All other ions will be subjected to ever increasing horizontal or vertical amplitudes and will finally be ejected from the quadrupole. The quadrupole can be configured to either select for a very narrow m/z range or to allow a broad range of m/z values to pass.
The structure of a quadrupole is depicted in figure 4.

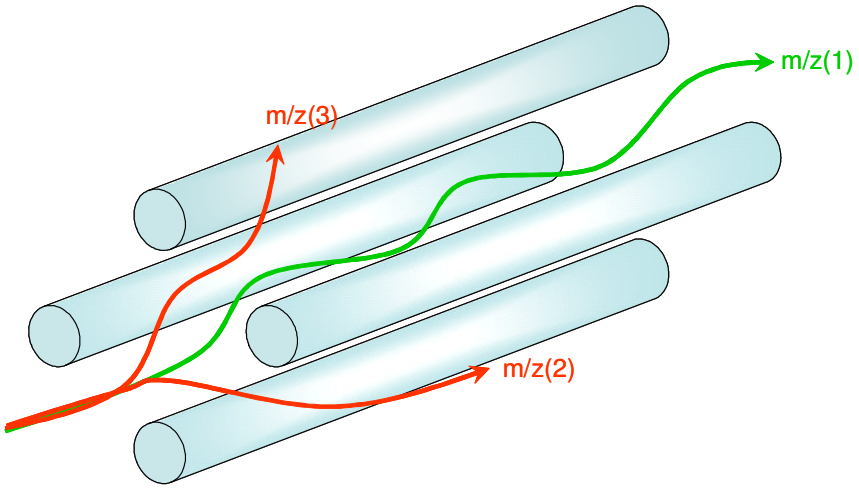


Figure 4: Schematic representation of a quadrupole mass analyzer.

The final mass analyzer discussed here is the ion trap mass spectrometer [March 1996, Jonscher 1997]. An ion trap consists of three electrodes: two capping electrodes at each end of the ion trap (one end is connected to the inlet from the ion source, the other leads to the detector) and a ring electrode that surrounds the trap chamber. The ions enter the trap and are subjected to an oscillating radiofrequency electrical field very similar to that employed in a quadrupole. This field causes the ions to move first back- and forward axially and inward radially, and then outward radially and inward axially, generating an ion cloud that is continually compressing and expanding along the axes. An inert cooling gas (typically helium) is added to the chamber to absorb excess energy from the ions through collisions. Since the voltage on the ring electrode determines the m/z ratio cut-off below which the ions are ejected out of the trap, stepping up the voltage on the ring electrode allows the ion trap analyzer to progressively scan across a mass range.
A schematic cross-section view on an ion trap is presented in figure 5.

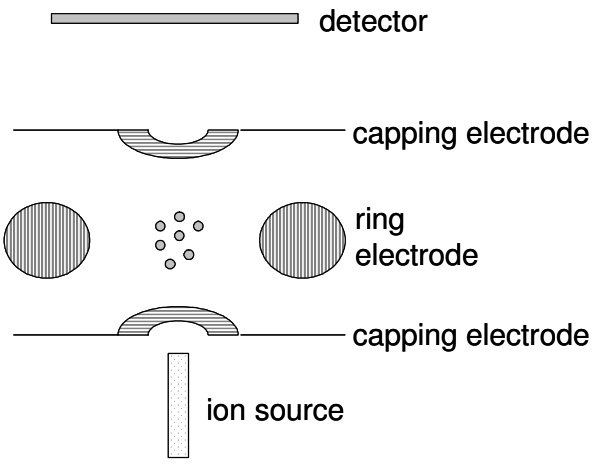


Figure 5: cross-sectional schematic view of an ion trap.

1.1.3.3. Fragmentation mechanisms

There are two important fragmentation mechanisms available in current-day mass spectrometers. The first of these is specific to MALDI sources and is called post-source decay (PSD). The second is called collision-induced dissociation (CID) and can be applied regardless of the source that produced the ions.
Ions derived from a MALDI source are typically highly energetic and thus metastable. A significant percentage of these ions spontaneously decays into fragments during the trip in the field-free vacuum tube. This process is thus unimolecular and the resulting fragment ions are indiscernible in a linear TOF as their velocities are identical to that of their parent ion. By employing a reflectron field, the changes in kinetic energy due to the smaller mass of these fragment ions can be used to separate them. This application of the reflectron field has been illustrated in figure 6.

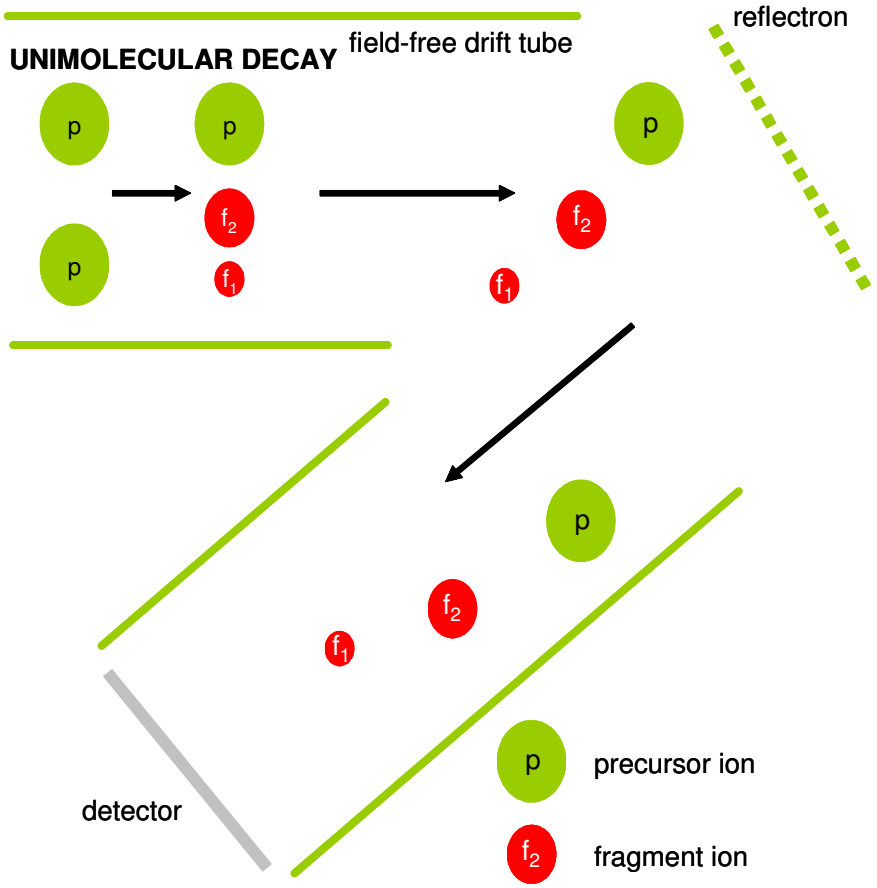


Figure 6: schematic representation of the separation of precursor and fragment ions after post-source decay using a reflectron field.

CID on the other hand is a bimolecular process in which accelerated ions are sent through a relatively dense collision gas (typically noble gasses, although nitrogen or even air are sometimes used as well). The random collisions between an analyte ion and collision gas molecules can then lead to the fragmentation of the ion. Important parameters that influence the fragmentation process are the pressure of the collision gas as well as the

collision voltage. The latter is therefore often ramped over time to explore as broad a range of energies as possible.
Fragmentation ions fall into three broad categories: those derived from the N-terminal end of the original peptide ion, those from the C-terminal end and internal fragments. Each of these fragments can be broken in several places along the peptide bond. The resulting fragments and their one-letter designation [Roepstorff 1984, Biemann 1988] are represented in figure 7.

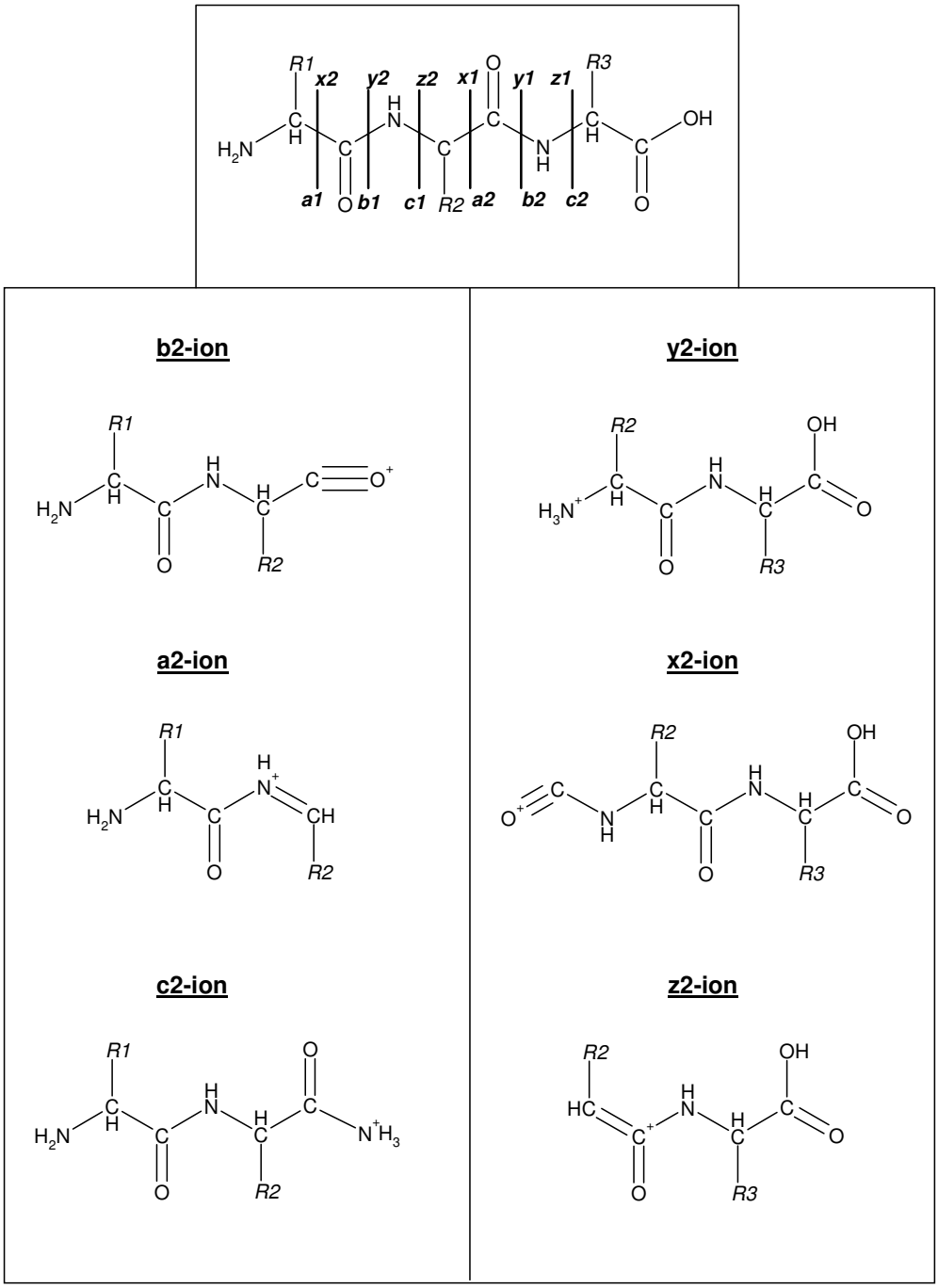


Figure 7: structure and nomenclature for the most common peptide fragment ions.

1.1.4. 2D-PAGE proteomics and where they fail

Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) [O'Farrell and Klose 1975] has been the workhorse of proteomics for over three decades. Its ability to separate complex protein mixtures in two orthogonal dimensions according to different physico-chemical properties was fortuitously supplemented by the advent of protein identification by mass spectrometry. The separation by 2D-PAGE relies on protein iso-electric point (pI) in the first dimension (isoelectric focusing, IEF) and apparent molecular weight (SDS-PAGE) in the second dimension.

In a typical analysis pipeline, a protein mixture is first separated on a 2D-PAGE system and the resulting proteome pattern is subsequently visualized using a staining protocol. Popular examples of these are Coomassie brilliant blue [Meyer 1965] and silver staining [Switzer 1979].

Visualized protein spots can then be excised followed by proteolytic digestion either after dissolving the proteins from the gel or, more generally, directly inside the gel itself (*in-gel* digestion). The proteolytic enzyme of choice is usually trypsin, whose propensity to cleave on the carboxy-terminal (C-terminal) side of arginine or lysine [Olsen 2004] generates peptides with at least one of these basic amino acids. This property is advantageous when making the step to mass spectrometry as mass spectrometers can only analyze charged ions.

Identification of the isolated protein(s) then proceeds through a matching of the peptide masses recorded by the mass spectrometers to the masses of the *in silico* generated cleavage products from the entries in a protein sequence database. This comparison is usually too cumbersome to perform manually and specialized software has been written to allow the automated matching and scoring of peptide masses to protein databases [Pappin 1993, Mann 1993, Yates 1993, Clauser 1999, Colinge J 2003, Geer 2004][3]. Such algorithms usually take a coverage factor into account when scoring a protein hit. This behaviour builds on the assumption that the protein digest generates full sequence coverage through the resulting peptides for the parent protein. Furthermore, it also expects the mass spectrometer to recover as many peptides as possible from the sample presented to it. Additionally, these algorithms can also adjust the score by employing the principle of completeness: all non-matching masses in a spectrum can be considered to provide evidence contrary to the hit that is being scored. This builds on the assumption that proteins are isolated in pure form after the 2D separation.

2D gel-based approaches have certainly proved their usefulness over the years, yet some fundamental problems with this approach also became evident over time.

One of the first problems noticed concerned basic proteins. The first dimension separates proteins based on pI by applying a voltage across a pH gradient in which the proteins are free to migrate. Proteins thus travel in the electric field towards the pole that has opposite charge compared to their own. During this migration the protein will encounter various pH buffering conditions and its net charge will be adjusted accordingly. Whenever a protein reaches the location where its pI equals the local pH, its net charge will be zero, immediately blinding the protein in question to the voltage gradient. At this moment, migration will cease and the protein will stay in this location. The problem with highly basic proteins is that the pH gradient cannot accommodate the extreme pI of these

[3] These algorithms are further discussed in section 1.2.1.

proteins. As such, these proteins will simply continue to migrate until they finally elute from the strip. Possible solutions for this problem have been proposed, but these tend to be very work-intensive and can only be applied to specific cases [O'Farrell 1977, Görg 1988, Görg 1997, Görg 1999].

A second consideration centers on low-abundance proteins. As the 2D-PAGE approach relies on protein staining to visualize proteins prior to excising them and submitting them to further analysis, the recovery of low-abundance proteins hinges on the resolving power of the staining procedure. It turns out that the even the best staining protocols can only resolve the most abundant protein spots [Gygi 2000]. A complementary problem is that in the IEF phase there is a limit to the amount of material that can be applied. As such, 2D-PAGE analyses lack the dynamic range required to gain a complete view on the proteome. Additionally, low-abundance proteins are quite often highly important and influential in the living cell and therefore highly interesting study subjects [Marko-Varga 2003].

Another class of proteins that presents specific challenges to 2D-gel based analyses are hydrophobic proteins. These proteins are difficult to fully and reproducibly extract from their natural milieu (i.e. hydrophobic biological membranes) and furthermore have the tendency to precipitate easily during the isoelectric focusing step, preventing them from penetrating into the second dimension and are thus lost. This problem has even proved frustrating enough to prompt lyrical review titles: 'Membrane proteins and proteomics: *un amour impossible?*' [Santoni 2000]. Typical examples of hydrophobic proteins are membrane proteins and, since these are crucial for a cell to explore and communicate with other cells and the outside world, not being able to detect them is a major impediment to achieving understanding of a living cell through proteomics.

1.1.5. Skipping the protein phase

Researchers started to tackle the problems inherent in 2D-PAGE proteomics by relying less on the protein as a unit for separation and identification, and more on the peptides that result after proteolytic digestion.

The term *peptide-centric*[4] proteomics is typically used to designate these technologies. This shift from proteins to peptides presents several benefits:

- distributions of properties are less broad: peptides tend to suffer less from extremes in their physico-chemical properties;
- peptides open up ways to employ a power-by-numbers strategy: a single protein can typically yield tens of tryptic peptides (the average protein size in SWISS-PROT release 48.7 is approximately 340 amino acids, with arginine residues comprising 5.35% of all amino acids in the database and lysine accounting for 5.93%). Where protein-based separation represented a binary situation (either the protein is recovered, or it is missed), the peptide-based separation can now be related to a more fuzzy situation (x out of y peptides are recovered for a given protein, versus (y-x) that are missed).

[4] Initially these technologies were catalogued as *gel-free* proteomics. Since this is a negative definition and therefore not very informative about the actual concept, it can be said that this was not the best of choices. *Peptide-centric* proteomics, as a positive definition, is much more descriptive and has therefore taken preference over the last years.

- peptides are in general easier to handle: e.g., they do not precipitate that readily and contemporary mass spectrometers have an inter-spectrum dynamic range of several orders of magnitude making it possible to simultaneously analyze abundant and scarce peptides.

It is obvious from the above that the proteolytic digestion of proteins into peptides before proteomic analysis introduces significant redundancy in the sample. This can be deemed both advantageous and problematic. An example of an advantageous effect is that large, hydrophobic proteins will yield quite a few hydrophobic peptides that remain difficult to separate and identify however, chances are very good that among the many peptides obtained after cleavage, at least some are more readily amenable to standard analysis techniques and that these are finally identified. The redundancy introduced by shifting to the peptide level thus enables more straightforward identification of the parent protein. Alternatively, having many distinct peptides instead of a single protein also increases sample complexity dramatically. This increase in complexity needs to be subsequently accommodated for in the separation phases.

1.1.6. Separation revisited

With the focus on the peptide instead of the protein, the separation methods employed necessarily needed to change as well. The first published protocols [Link 1999, Washburn 2001, Wolters 2001, Washburn 2002] were two-dimensional peptide-separation techniques in which the familiar dimensions of 2D-PAGE protein separation were each replaced by peptide equivalents. The iso-electric focusing for proteins was replaced by ion-exchange chromatography and the MW separation was replaced by reverse-phase HPLC. This technique is called Multidimensional Proteome Identification Technique (MudPIT) and is illustrated in figure 8.

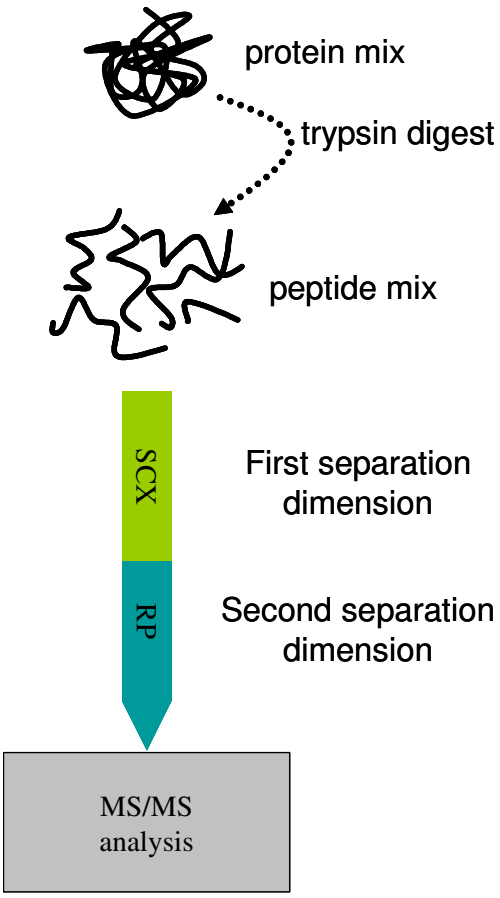


Figure 8: Principle of the MudPIT technology.

1.1.7. Selection as an alternative to separation

As calculated above, the digestion of proteins by trypsin (by far the most commonly used proteolytic enzyme in proteomics) yields on average thirty, ten amino acid long peptides per protein. Assuming these peptide sequences are all unique, this directly leads to a dramatic increase in sample complexity. The separation methods employed in peptide-centric proteomics however, cannot accommodate such a saturated mixture even after separating peptides in two dimensions. Consequently, the mass spectrometers analysing the resulting badly separated sample will be unable to select every passing peptide for fragmentation [Liu 2004].
For this reason, techniques have been developed that isolate representative subsets from the peptide mixture in order to reduce its complexity rather than relying solely on separation to spread the sample constituents over time [Gygi 1999, Spahr 2000, Wang & Regnier 2001, Oda 2001]. These methods are briefly discussed below:

- *Isotope coded affinity tag* (ICAT) [Gygi 1999].
  The original ICAT molecule selects for cysteine-containing peptides by exploiting the highly reactive properties of the sulfhydryl side-chain of the cysteine. Typically, an affinity label (biotin) is attached to this side-chain via a thiol-reactive group such as iodoacetamide. A linker segment between these two groups is also present, which

can carry up to eight ^{2}H (D) residues for differential labeling. Subsequently, cysteinyl peptides are affinity-isolated (streptavidin) and analyzed by MS techniques. The ICAT molecule is shown in figure 9.

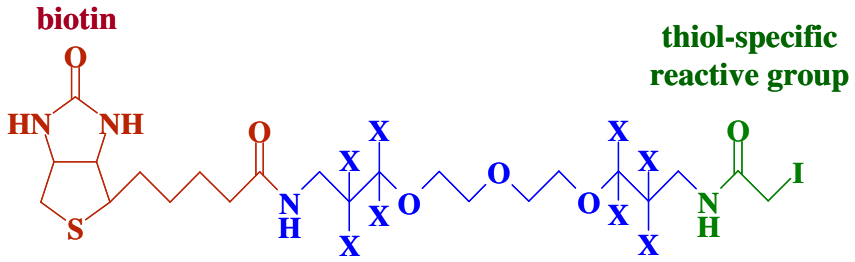


Figure 9: The ICAT molecule. Note the location of 'X' residues. These locations can contain either ^{1}H or ^{2}H (D), allowing the labeling strategy to be used for relative quantification of proteins.

The principle of tagging peptide subsets in order to enable their subsequent isolation has been extended to include capture of diverse sequence elements and peptide modification states [Oda 2001, Wang 2002, Zhang 2003, Peng 2003, Khidekel 2004, Kho 2004, Denison 2005].

- *Combined fractional diagonal chromatography* (COFRADIC) [Gevaert 2002, Gevaert 2003, Gevaert 2004, Gevaert 2005].
  COFRADIC is a technique based on diagonal chromatography of peptides in a HPLC system. The technique employs two identical reverse-phase (RP) HPLC runs on the same system, introducing a modification step in between the two runs that selectively targets certain peptides in the mixture by altering their chromatographic properties. Peptides can now be isolated because they show changed chromatographic behaviour ('forward' COFRADIC), or because they retain their original behaviour while all other peptides shift ('reverse' COFRADIC). A general depiction of the COFRADIC sorting procedure is given in figure 10. COFRADIC is discussed in detail in the next section.

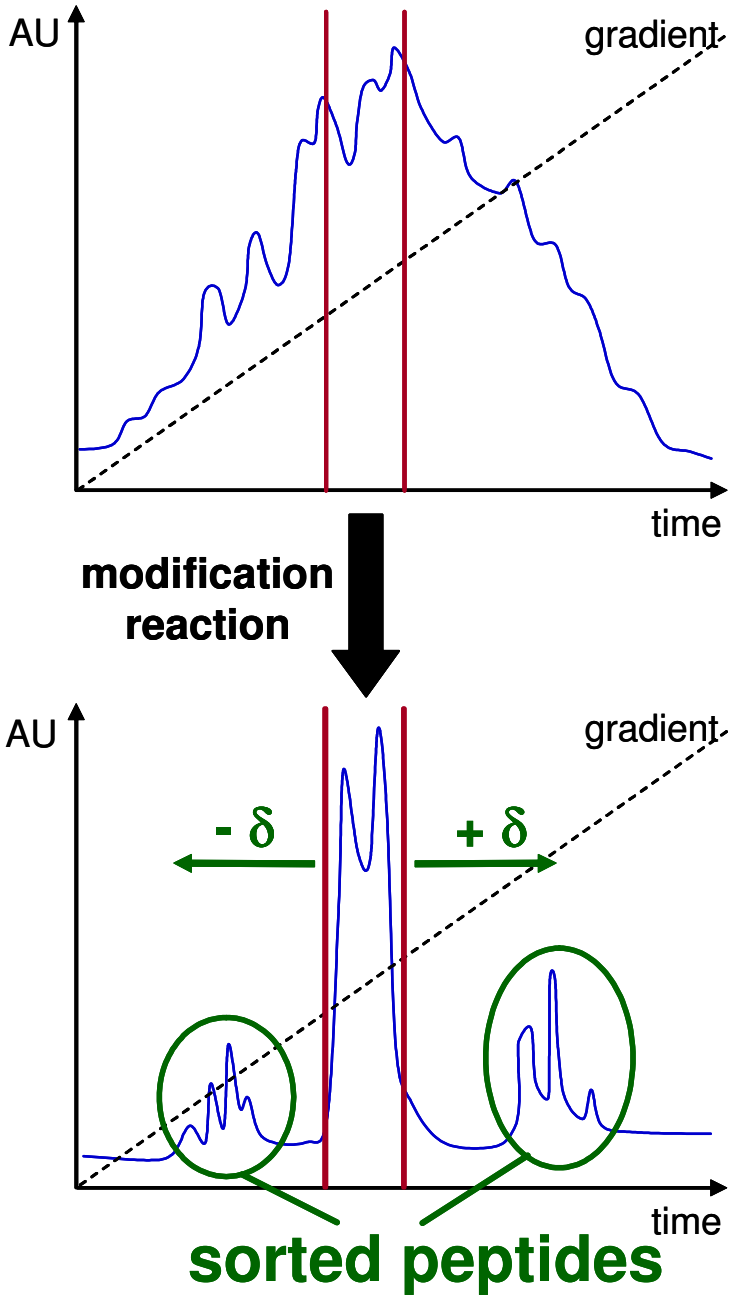


Figure 10: General COFRADIC sorting principle. The changed chromatographic properties of targeted peptides following the modification reaction moves them outside of the original elution fraction (lower chromatogram).

These techniques all have to maximize the reduction of sample complexity whilst simultaneously minimizing the amount of proteins that are lost due to the absence of identifiable peptides in their sequences. It is clear that the selection procedure is often based on sequence features of the peptides and that the most popular targets are therefore modifiable rare amino acids and functional groups. The relative scarceness of these amino acids allows them to act as prime complexity reducers. As an illustration, the theoretical reduction in complexity as well as the loss of sample coverage against the

SWISS-PROT database (dd. 24th of January 2006) are given for the most popular rare amino acids and proteolytic enzymes in table 1 below.

Enzyme	Selection	Complexity reduction	Proteins missed
Trypsin	*All*	0%	0.16%
	Met	74%	1.73%
	Mu	77%	4.34%
	Cys	75%	4.22%
	His	70%	3.78%
	Trp	84%	9.29%
Arg-C	*All*	0%	0.95%
	Met	71%	4.97%
	Mu	75%	9.95%
	Cys	73%	9.61%
	His	66%	8.43%
	Trp	82%	17.75%

Table 1: Theoretical complexity reduction and unrecoverable protein fraction using several common subset isolation strategies for both trypsin and ArgC digests of the human subset of SWISS-PROT (24 January 2006). 'Mu' denotes methionine residues minus initiator methionines.

1.1.8. COFRADIC

1.1.8.1. Methionine COFRADIC

Isolation of methionyl peptides hinges on the modification reaction of methionine to its sulfoxide by peroxide [Gevaert 2002]. This reaction is shown in figure 11.

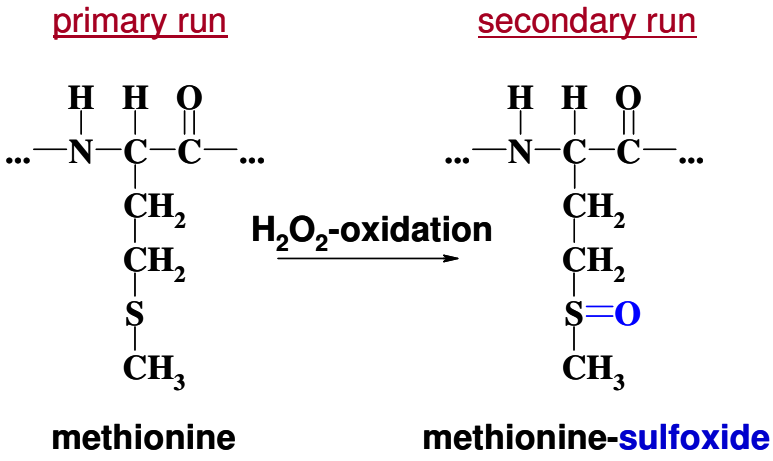


Figure 11: Conversion of methionine to its sulfoxide by peroxide between the primary and secondary COFRADIC runs.

Because of the dipole moment introduced, this methionine-sulfoxide is more hydrophilic than the original methionine, inducing a hydrophilic shift for all methionyl peptides. In order to optimize the throughput of this technique, several primary fractions can be pooled[5] prior to the secondary run reducing the number of required secondary HPLC separations. The pooling of four fractions is illustrated in figure 12.

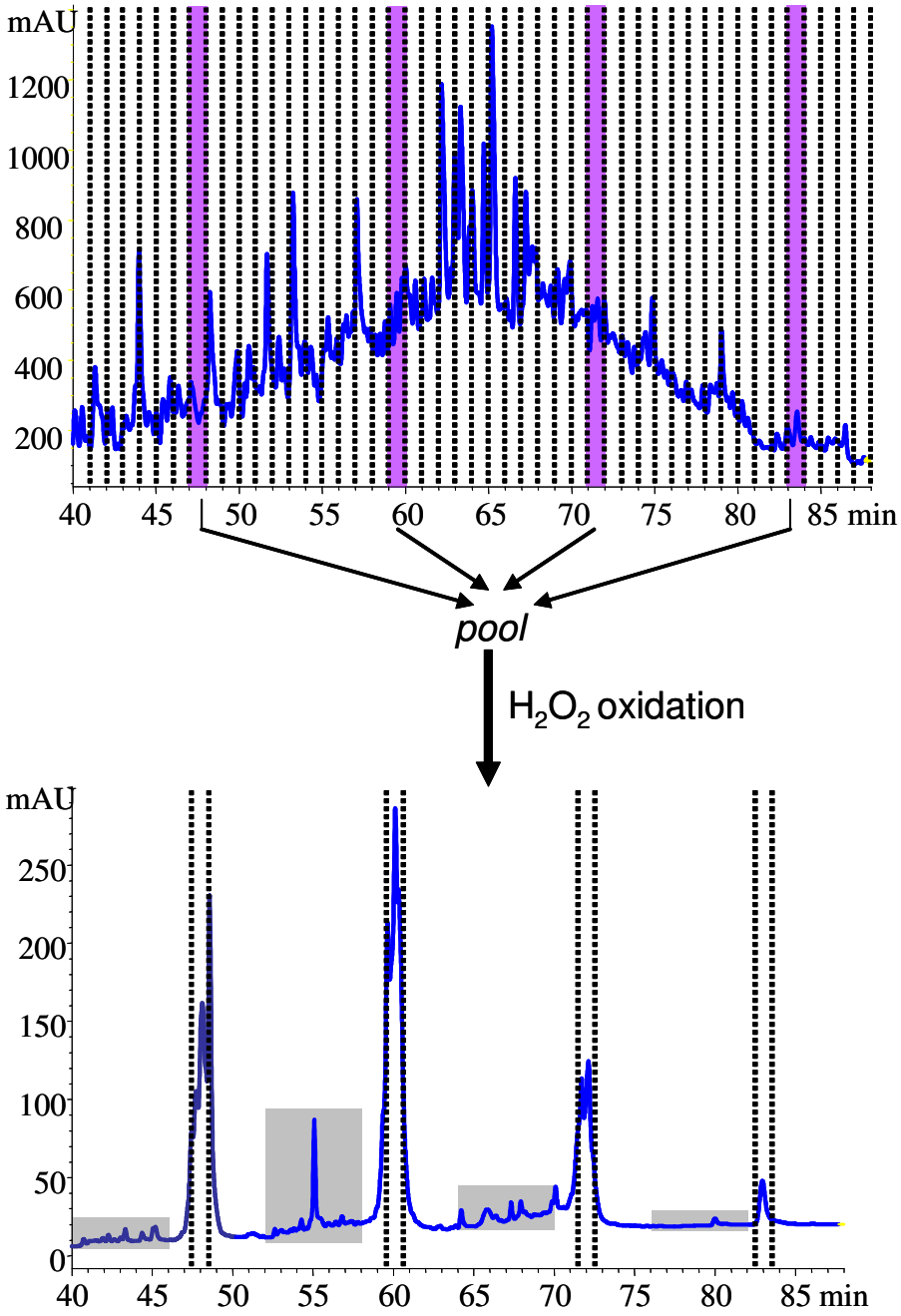


Figure 12: Pooling of four different methionine COFRADIC fractions. Traces of shifted peptides are set against a grey background in the second RP-HPLC step.

[5] Or 'combined', yielding the 'CO' in the acronym COFRADIC.

1.1.8.2. Cysteine COFRADIC

Isolation of cysteinyl peptides follows a slightly different scenario [Gevaert 2004]. The chromatographic shift is accomplished here by the removal of a thionitrobenzoyl (TNB) moiety from the cysteine side-chain between the first and second chromatographic step. The modification of cysteine residues with Ellman's reagent (5,5-dithiobis(2-nitrobenzoic acid) or DTNB) to obtain the mixed disulfide (TNB-cysteine) therefore takes place before the primary HPLC run. The TNB-cysteine is more hydrophobic than free cysteine and removal of the TNB moiety by tris(2-carboxyethyl)phosphine (TCEP) reduction prior to the secondary COFRADIC run results in a hydrophilic shift for cysteine-containing peptides. The entire process is represented in figure 13.

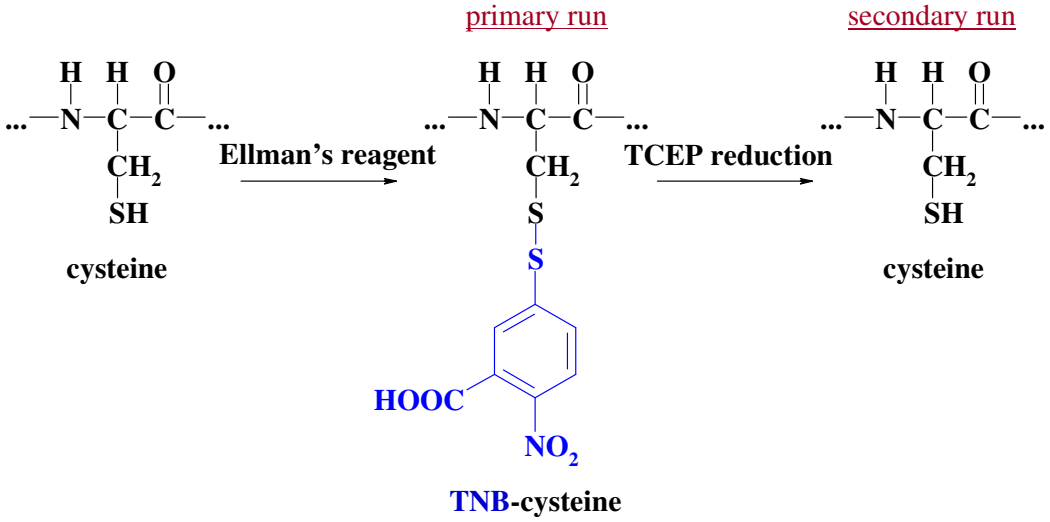


Figure 13: Cysteine modification with Ellman's reagent prior to the primary COFRADIC run yields TNB-cysteine which is more hydrophobic than regular cysteine. After the primary run, a simple reduction step decouples the thionitrobenzoyl moiety and the resulting free cysteines display a hydrophilic shift in the secondary run.

1.1.8.3. Amino terminal (N-terminal) COFRADIC

N-terminal COFRADIC is an odd duck in the pond when one considers peptide-isolation strategies [Gevaert 2003]. Unlike most other strategies, this technique isolates peptides based on a positional parameter instead of their amino acid composition. Another peculiarity of N-terminal COFRADIC is that it is considered 'reverse' COFRADIC since the chemistry employed between the primary and secondary separation step will shift all non-N-terminal peptides. The collected peptides will therefore be the non-shifting ones that have retained their original chromatographic properties.

In practice, the sample is first prepared by the addition of iodoacetamide to alkylate free cysteines and subsequently treated with sulfo-N-hydroxysuccinimide acetate to

acetylate all free amines (both α- and ε-amines). Note that up to this point, all reactions are applied to a protein mixture. The next step is tryptic digestion. Because the ε-amines of the lysine residues have been acetylated, trypsin will only cleave C-terminally of arginine residues, thus effectively behaving like endoprotease Arg-C. There are now two kinds of peptides in the mixture: those that formed the original protein N-termini and those resulting from the proteolytic cleavage with trypsin. The former carry acetylated α-amines while the latter present a free α-amine. The next step then is to perform the primary COFRADIC run. After this first run, the collected fractions are subjected to 2,4,6-trinitrobenzenesulfonic acid (TNBS). This reagent reacts with free amines to form trinitrophenyl-peptides. Since only the internal peptides possess free amines, these are the only ones modified. The trinitrophenyl-peptides are much more hydrophobic than their original counterparts and will therefore shift outside the original collection interval. Since the targeted peptides are the unaltered N-terminal peptides the non-shifting peptides will be collected here. The TNBS reaction step is displayed in figure 14 for both N-terminal peptides and internal peptides.

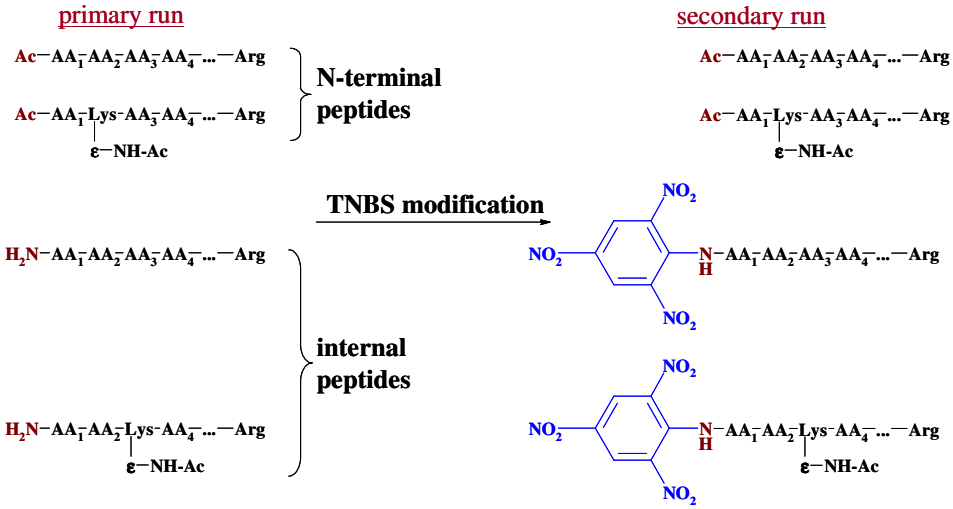


Figure 14: TNBS modification reaction. Note that cysteines have been alkylated and protein N-termini and lysines have been acetylated.

Example chromatograms (UV absorbance measured at 214 nm) for the primary and secondary run are shown in figure 15, illustrating the concept of 'reverse' COFRADIC.

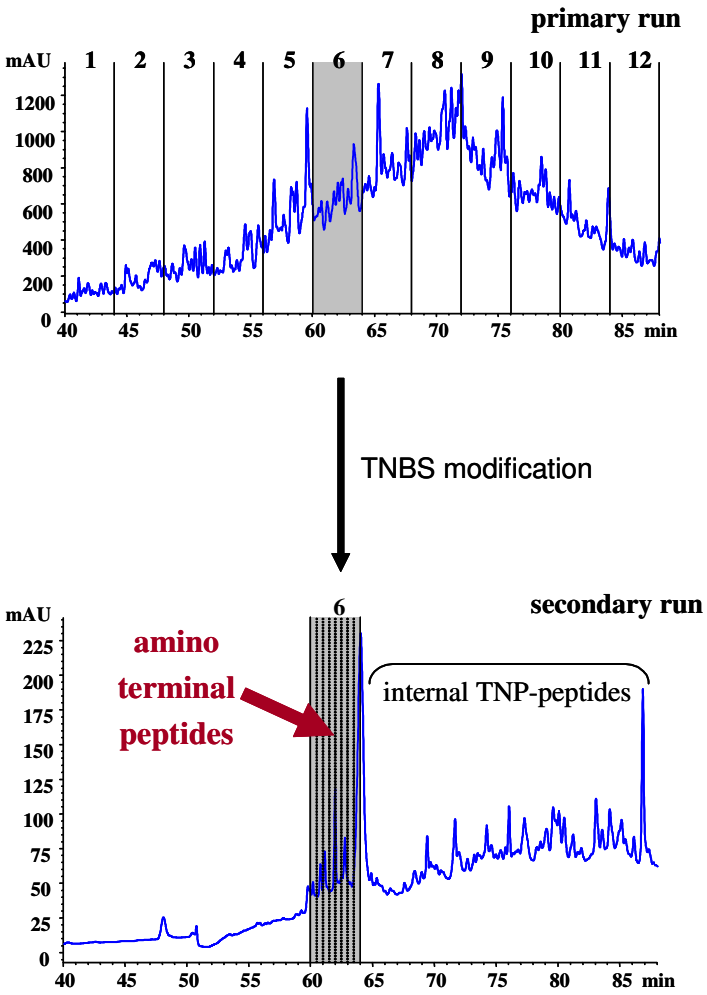


Figure 15: N-terminal or 'reversed' COFRADIC. Note that the shifting peptides are the unwanted fraction here.

1.2. Identification of proteins by mass spectrometry

The process by which mass spectrometry data is converted into protein lists is governed by two important assets: software algorithms that take care of identifying mass spectra and, for most of these, the protein sequence databases they use as search space.
The various types of identification algorithms as well as the importance and characteristics of protein sequence databases will be discussed in the next subsections.

1.2.1. Identification algorithms

1.2.1.1. Classification and description of common approaches

Identification algorithms can be roughly divided in two separate groups: those that utilize solely the information contained in a spectrum (*de novo* algorithms) and those that require a sequence database to search against (search algorithms).
The former approach is often used when unknown post-translational modifications need to be picked up or when there is insufficient proteomic or genomic information available for the organism under study to apply search algorithms. De novo algorithms usually require very clean fragmentation spectra and high ion series coverage in order to operate reliably. When applied to high-throughput spectra this usually leads to incomplete sequencing or reporting of many indiscernible possibilities. Because of the problems associated with coupling de novo sequence analysis to high-throughput proteomics and the growing availability of good proteomic or genomic sequences databases for nearly all model organisms, these identification algorithms tend to be reserved for semi-automatic or manual application in special cases. Several algorithms exist, including Lutefisk, Sherenga, PEAKS and PepNovo [Dancik 1999, Taylor 2000, Fernandez-de-Cossio 2000, Ma 2003, Zhang 2004, Frank 2005, Grossmann 2005].
The search algorithms all attempt to match a spectrum (or spectrum-derived information) to sequence databases at some point during their processing. There are two main approaches to matching a spectrum to a database, the first of which represents a hybrid between de novo and search algorithms. A spectrum is first examined for the presence of short sequence tags, small groups of a few consecutive amino acids that can be extracted even from spectra with incomplete ion series coverage. These sequence tags are subsequently matched against the database. Several tag-based engines have been developed over time, with PeptideSearch [Mann 1994] being one of the first and GutenTag [Tabb 2003] one of the most recent algorithms. A completely different algorithmic approach with a similar mixture of de novo sequencing and database searching is taken by Popitam [Hernandez 2003]. This software uses an Ant Colony Optimization algorithm to attempt the fitting of sequences to a spectrum in a massively parallel way. The result is a set of ‘walks’ through the spectrum, each of which is inspired and limited by a sequence from a database. When a walk is highly successful, it will become more pronounced, attracting more ‘ants’ to further explore it. The end result is an optimal explanation of the spectrum by a sequence in the database.
The other type of database search algorithm is fully database-centric. In this category the two most popular search engines in the field today can be found: Mascot [Perkins 1999] and SEQUEST [Eng 1994]. Apart from these two commercial algorithms, some open

source search engines are also available, most notably OMSSA [Geer 2004] and X!Tandem [Fenyo 2003]. All of the algorithms in these categories first generate *in silico* peptide sequences from the available protein sequences and then construct theoretical fragmentation spectra from these peptide sequences. These *in silico* spectra are then compared with the experimental fragmentation spectrum and a score is assigned to the match. The difference between these algorithms can be found in the scoring function applied and whether or not thresholds are reported.
SEQUEST applies a cross-correlation function to derive its main score (a preliminary, filtering score called 'p-score' is first calculated based on the number of matching fragment peaks, possible sequence continuity in them and their intensities) where Mascot uses the MOWSE score function [Pappin 1993]. Another main difference between SEQUEST and Mascot lies in the fact that Mascot also provides a statistically relevant threshold score for each spectrum. Whenever a peptide match scores below this threshold, the identification should be disregarded. This threshold is set at a 95% confidence interval by default, but can be readily adapted in stringency if desired. Even though SEQUEST does not have a built-in thresholding scheme, the Expectation-Maximization and Bayesian statistics applied post-identification by the PeptideProphet algorithm [Keller 2002] can provide a highly useful distinction between true positives and false positives for SEQUEST. It is of note that the combination SEQUEST/PeptideProphet achieved very similar results to Mascot in a large-scale test performed by Kapp *et al.* and that these two algorithms clearly outperformed the other algorithms tested [Kapp 2005]. Other approaches to achieve a better discrimination between true and false positives for SEQUEST include RScore [Li 2004] and machine-learning strategies [Anderson 2003, Baczek 2004].
The approach taken by the OLAV [Colinge 2003] algorithm refines the generation of a theoretical fragmentation spectrum from a sequence in a database by a stochastic prediction of the chances for each fragment ion to appear, thus attempting to predict the actual measured spectrum as well as possible before scoring the match.

1.2.1.2. Mascot threshold calculation

Since the work described in this dissertation has been performed using the Mascot search engine, the threshold calculation performed by this particular algorithm is examined in a little more detail here. It is important to know that the actual MS/MS-based identification engine of Mascot has not been published, and that all information about its workings has therefore to be inferred from its observed behaviour. Within the scope of this text however, these indirect clues are more than sufficient.
A probabilistic algorithm such as Mascot attempts to estimate an *a priori* chance of random matching against a sequence database. This estimate necessarily relies on the number of unique peptide sequences in the search base, which we will here call the *information content* of the database for a specific proteolytic enzyme. In order to provide a practical estimate of this information content for a certain proteolytic enzyme, Mascot takes the size of the database (expressed as the total number of amino acids in that database) and the prevalence of the recognized residues for cleavage into account when calculating the identity threshold score. Indeed, the average identity threshold score for a certain enzyme in a certain sequence database is the reciprocal of the logarithm to base

10 of the total number of amino acids in that database. This relationship is illustrated for trypsin across a few often-used databases in figure 16.

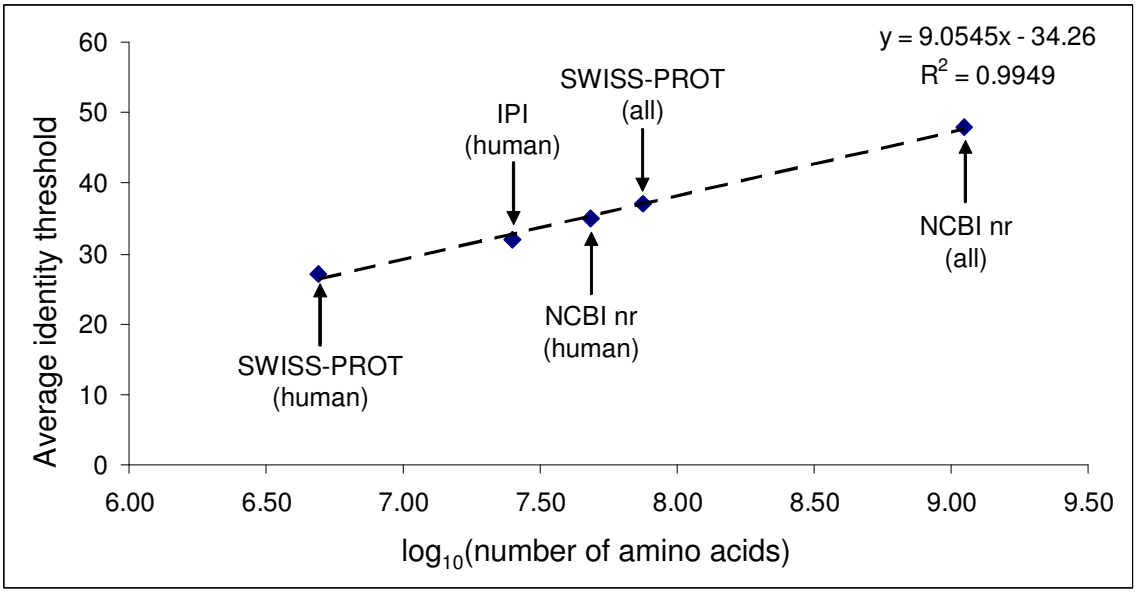


Figure 16: Relationship between the number of amino acids in a search space and its average identity threshold score for a trypsin digest in Mascot.

1.2.2. Protein sequence databases

As has been outlined in section 1.2.1.1, high-throughput protein identification is usually based on database search algorithms. Interestingly, even though the sequence database employed represents the most basic source of information in these identifications, its importance is often disregarded. In the next three subsections the most commonly used sequence databases for protein identification will be briefly discussed and in the fourth subsection a note is given on the instability in time of sequence databases.

1.2.2.1. UniProt KnowledgeBase (SWISS-PROT/TrEMBL)

The central UniProt [Wu 2006] database builds on two well-established pillars in the sequence database world: the SWISS-PROT and TrEMBL databases. The key difference between SWISS-PROT and TrEMBL[6] lies in the manual curation effort that underlies the former. Indeed, all entries in SWISS-PROT have passed through a rigorous manual control by human curators. During this curation process diverse information sources are consulted and cross-verified in order to establish which annotations are clearly supported by trustworthy evidence. The result of these efforts is an extremely high-quality and stable[7] protein sequence database with heavily cross-linked annotations. Obviously, the

[6] This statement can actually be extended to 'the key difference between SWISS-PROT and any other popular protein sequence database'.

[7] The number of sequences curated into SWISS-PROT continues to grow, so 'stable' here simply means that it is unlikely that an entry will be deleted from SWISS-PROT.

curation process is labour-intensive and this limits the reach of the SWISS-PROT database. The TrEMBL database complements this nicely, however. It contains automatically annotated proteins, including predicted protein sequences for which corroborating evidence exists.
UniProt is more than simply a list of sequences and their annotations, however. Apart from the UniProt KnowledgeBase, the system encompasses a complete sequence archive called UniParc and UniRef, a set of reference clusters at different sequence identity levels.

1.2.2.2. The NCBI non-redundant database

The National Center for Bioinformatics Information (NCBI, http://www.ncbi.nih.gov) provides a non-redundant sequence database that is usually referred to as the 'NCBI nr' database. This database groups sequence information from a variety of sources, including SWISS-PROT, TrEMBL and RefSeq. The latter consists of two distinct types of entries, NP and XP, which are readily identifiable by their accession numbers. The NP sequence have corroborating evidence such as cDNA to back up their validity while the XP sequences are based purely on predictions. It is clear that the level of annotation and available cross-links varies considerably between entries and is largely dependent on the source of the sequence.
The database is non-redundant at the absolute protein sequence level, meaning that no two sequences are completely identical in the database. History management is also provided via the Entrez web interface (http://www.ncbi.nlm.nih.gov/entrez/).

1.2.2.3. The International Protein Index (IPI)

This database finds its origins in the human genome project [Lander 2001] and was originally conceived as a non-redundant view on all known human proteins. Over the years the reach and scope of the IPI database has grown [Kersey 2004], yet the basic premise of providing an intelligently designed non-redundant protein sequence database has remained. IPI is now available for a variety of model organisms, including human, mouse, rat and *Arabidopsis*. IPI now presents an automatically curated view on the total contents of a large collection of sequence databases (including UniProt, RefSeq and EnsEMBL) by a rather complex algorithm to remove sequence redundancy in a thorough way. Simply put, the algorithm can be described as follows: instead of simply requiring each protein sequence to be unique, sequences are also collapsed into clusters when they show more than 95% overlap over the full match length. When a cluster consists of sequences of different lengths (which often happens due to the presence of protein fragments as separate entries in several source sequence databases), the longest sequence is chosen as the master sequence. Notable exceptions to the clustering rule are the annotated UniProt splice-variants, which are retained as separate entries. Each cluster is finally assigned an IPI accession number and all source references aggregated in the cluster are expressly reported in the IPI entry.
IPI also provides complete history files, which trace the history of every entry that has ever carried an IPI identifier.

1.2.2.4. Time-instability of sequence databases

In the above sections it has been mentioned that each of the sequence databases maintains sequence history in a specific way. The ability to trace a sequence (or its (versioned) accession number) through time is once again a much-overlooked yet highly important characteristic of a sequence database.

Indeed, sequences in any database are subject, to a greater or lesser degree, to changes over time. The primary event of any sequence is of course its first inclusion in a database. This can be traced for each entry in the abovementioned sequence databases. Once a sequence has been recorded in a database, it can undergo alterations as well. Depending on the database, these alteration events can trigger a change in the accession versionnumber or even cause the assignment of a new accession number. In the latter case, tracing is usually provided in such a way that the original accession number will still link up to the sequence.

Certain sequences can be subjected to removal from the database, invalidating the accession number without providing a replacement. This effect is most pronounced for purely theoretical predictions (such as the abovementioned RefSeq XP entries). Indeed, a change in the prediction algorithm usually renders a number of previous predictions obsolete. From figure 17 it can be clearly seen that this deletion of spurious predictions has been most clearly noticeable over the period between December 2001 and December 2003, inclusive, for RefSeq XP. Also note that a database such as IPI, which combines information from different sources (including RefSeq XP), suffers the same fluctuations as its source databases.

An often-heard complaint from proteomics researchers concerns this deletion of sequences that have been identified by proteomics approaches. For example, using the IPI version of April 2002 one can clearly and uniquely identify several proteins that are derived solely from RefSeq XP. After a few months (for example in September 2002) re-examination of the data set may reveal that all RefSeq XP sequences have been deleted from the database. This issue is all the more unfortunate since the main difference between RefSeq XP and NP numbers lies in the availability of external evidence for the correctness of the prediction, which has in fact been furnished by the mass spectrometric identification of the sequence.

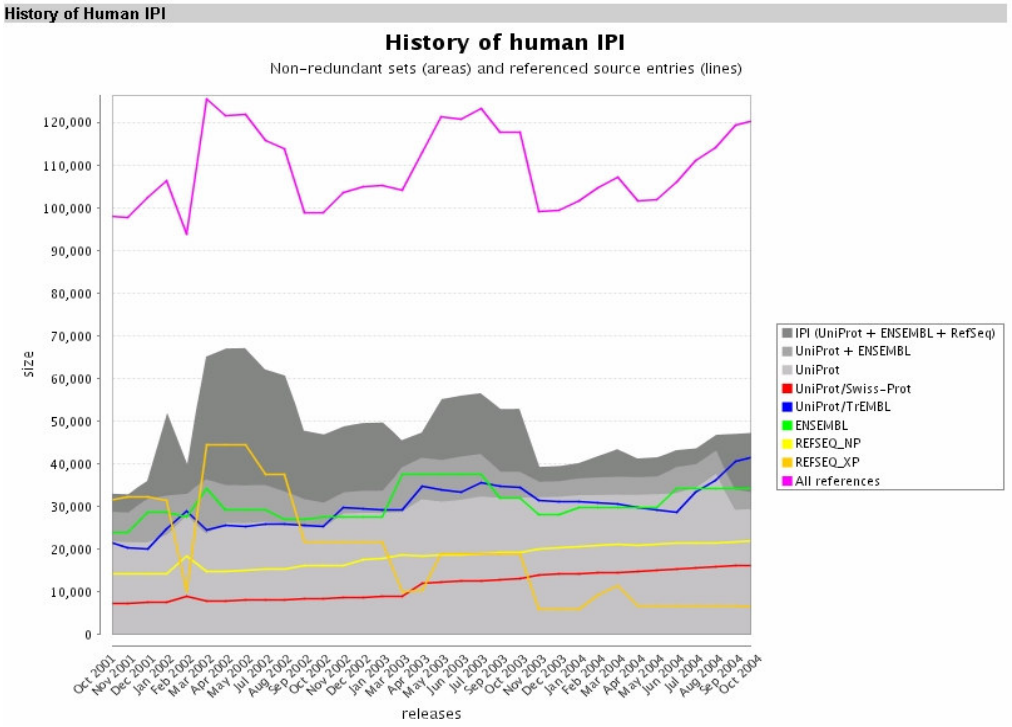


Figure 17: Changes in the IPI database and its subsidiary databases over the October 2001 – October 2004 period. Figure was obtained from the IPI website (http://www.ebi.ac.uk/ipi).

1.3. Mass spectrometry for the (relative) quantification of proteins

Proteomics studies often require differential labeling of two distinct samples to find proteins (or their modified forms) that differ in concentration. Some reviews are given here for differential labeling [Monteoliva 2004, Julka 2005] as well as for absolute quantification [Bronstrup 2004, Kirkpatrick 2005], although the latter falls outside the scope of this dissertation. In some cases, differential labeling of differently treated aliquots of the same sample may be required [Gevaert 2005]. The result of either approach can be seen as two distinct containers, each containing either a 'light' label or a 'heavy' label. This nomenclature is derived from stable isotopes such as ^{1}H and D, ^{16}O and ^{18}O, ^{14}N and ^{15}N or ^{12}C and ^{13}C. These isotopes differ only in their mass and thus yield otherwise identical compounds upon incorporation into molecules[8]. As mass is the only difference between the molecules subjected to differential isotopic labeling, it makes them ideal subjects for mass spectrometry where they can be readily separated into light and heavy peptides[9]. Subsequent analysis of the intensities of the corresponding m/z signals can yield insight into the ratio of these peptides in the mixture, thus allowing the experimentalist to obtain differential data and back-couple this to the state of the proteome.

The detectors utilized by mass spectrometers have their limitations in terms of dynamic range, however. This dynamic range was studied in detail for the detector of the Bruker Ultraflex instrument. A complete N-terminal COFRADIC sample was divided to two aliquots which were labelled with ^{16}O and ^{18}O respectively. The labelled peptides were subsequently mixed in different ratios, yielding eighteen aliquots ranging from a 1/9 to a 10/1 light over heavy ratio. Each of these mixtures was analyzed by the mass spectrometer and the resulting observed ratios were plotted against the known ratios in figure 18.

[8] In practice, elution characteristics on HPLC columns sometimes differ slightly for different isotopes of the same element, an effect that can become noticeable when incorporation of multiple isotopes (especially H/D) is used.

[9] Heavy isotopes occur naturally and will be incorporated into any biomolecules in ratios corresponding to their prevalence. As such, each biomolecule (such as a peptide) will display an isotopic pattern when analyzed by mass spectrometry. Certain incorporations such as single incorporations of ^{18}O or up to two incorporations of ^{2}H can lead to clearly overlapping isotopic patterns, presenting a more complex challenge in extracting the original abundances of light and heavy labeled peptides.

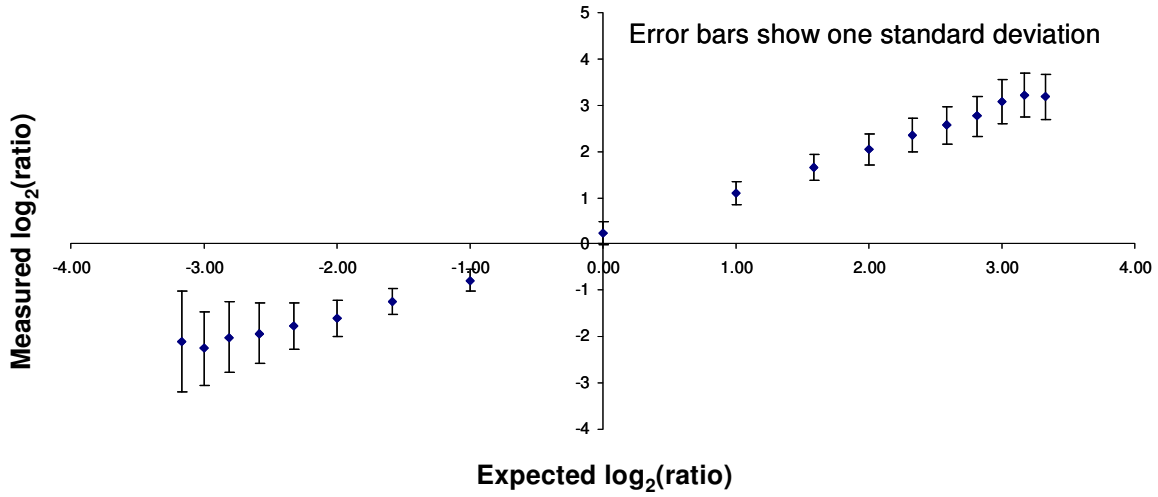


Figure 18: Dynamic range of the detector of the Bruker Ultraflex instrument.

The measured ratio is seen to level off at the extremes with the more extreme ratios also suffering from a larger standard deviation. There are two complementary explanations for this behaviour. The first one is that the detector response levels off at a certain amount of ions detected. This is due to saturation effects on the electron multiplier that amplifies detected signals. This saturation effect will be influenced by the actual amount of ions detected by the detector, which will depend on the amount of peptides spotted, the ionization properties of the peptides and those of other peptides in the same spot on the target[10].

The second effect is associated with the software processing of the mass spectra as recorded by the instrument. In order to calculate ratios, the instrument software must first determine peak intensities for both light and heavy isotopes. This in turn hinges on being able to detect a set of m/z peaks as the isotope envelope of a peptide, and this for both labeled peptides. When the ratios of mixed peptides approach a tenfold overabundance of either light or heavy labeled peptides, the signals generated by the minor component will be more readily lost in the noise, making detection difficult. This decreases the number of available measurements, negatively influencing the standard deviation.

[10] All peptides on the same spot compete with each other for the available charges. If a particular peptide is for instance much more prone to capturing a charge than all others, the resulting peptide ions will be prominently observed by the detector, even if this particular peptide is not dominantly present in the mixture contained within the spot.

1.4. Making every peptide count

1.4.1. Defining coverage

The term 'coverage' suffers from ambiguity in the context of this text and it therefore stands to reason that we clearly define it here. The number of proteins correctly identified from a mixture will be called 'sample coverage'. The coverage of a protein sequence by the individual sequences of the identified peptides will be referred to as 'protein sequence coverage'[11].

1.4.2. Concerning coverage and complexity

As has been explained in section 1.1.4, protein sequence coverage plays a significant role in the processing of identifications after mass spectrometric analysis of the spots on 2D gels. For the gel-free approaches however, this situation changes. Whereas MudPIT approaches (at least in theory) retain the ability to generate full protein sequence coverage, most of the peptide-centric techniques apply a peptide-isolation step to reduce complexity. Obviously, selecting for only a subset of peptides removes the possibility of having 100% protein sequence coverage. The extreme case of this is found in amino terminal COFRADIC, which (theoretically) yields only a single amino terminal peptide and obvious exceptions (e.g. peptides with TNBS inaccessible N-termini such as proline and pyroglutamic acid).
Since mass spectrometers often have difficulties fully analysing very complex samples [Liu 2004], gel-free techniques that retain a high level of protein sequence coverage (and therefore a highly complex peptide mixture) suffer from a corresponding loss of sample coverage. Approaches exist that can partially alleviate this problem under certain circumstances. These are detailed below:

- Increase separation [Wang 2005]
  By performing additional (preferably independent) separation steps the peptides are spread over a much broader range and the time between the presentation of individual peptides to the mass spectrometer increases. There are two major difficulties with this approach. The first of these concerns the additional manipulation of the sample. Due to surface adsorption, incomplete chemistry or precipitation effects losses occur, making it likely that low abundance proteins or their peptides have altogether disappeared from the sample by the time it reaches the mass spectrometer. Secondly, the increased separation steps can be laborious work, requiring skilled operators. Additionally, if a step n yields F_n fractions after separation, the number of times the $(n+1)^{th}$ step needs to be applied is given by the following multiplicative series:

[11] Note that protein sequence coverage can be exactly calculated when the peptide sequence and enzymatic cleavage applied are known, whereas the sample coverage is usually unknowable. Indeed, if we would *a priori* know the total number of proteins in a mixture, we would have already detected all of them, obviating the need for re-analysis.

$$\prod_{i=0}^{n} F_i := F_0 \cdot F_1 \cdot F_2 \cdot \ldots \cdot F_{n-1} \cdot F_n$$

This is also illustrated in figure 19.

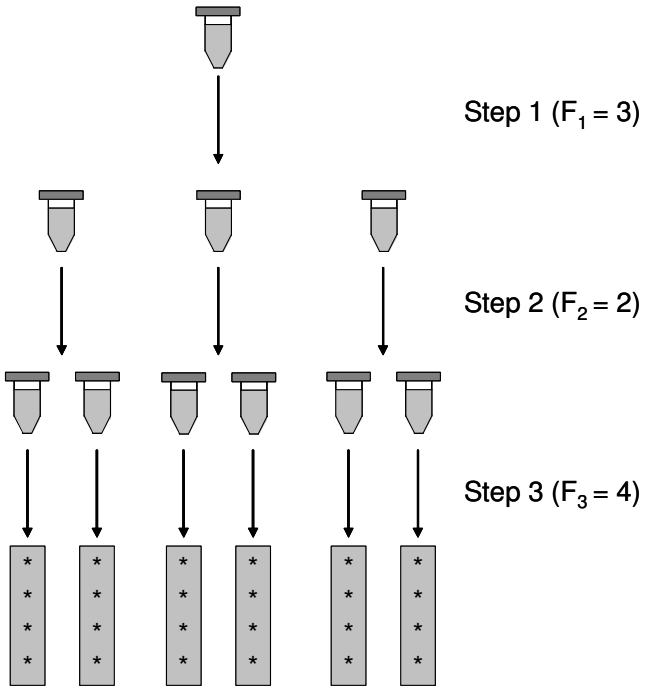


Figure 19: multiplicative increase in the number of separation steps required

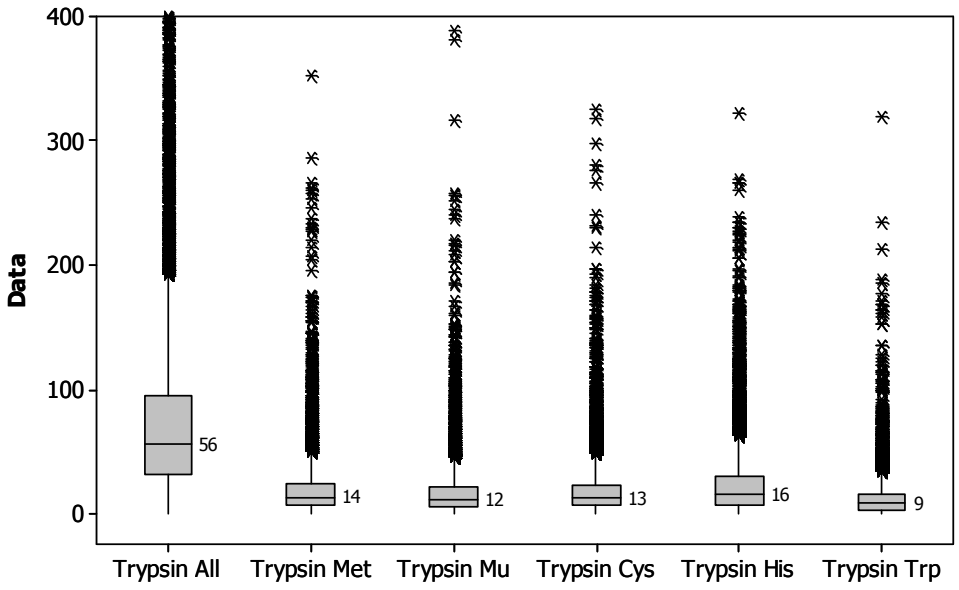
In all, the combined workload together with the poor automation of these steps and the finite time required to perform each separation make this system largely unsuited to high-throughput analysis.

- Use of dynamic (in time) and static exclusion lists [Spahr 2000]
  One can instruct most modern mass spectrometers with an ESI source to disregard certain peptide m/z values when selecting precursors to fragment during analysis. This can be done using both dynamic and static techniques. The dynamic exclusion approach is applied on the fly during the analysis: after having selected a particular precursor for fragmentation, the mass spectrometer will ignore that precursor mass for a preset time interval, enabling it to fragment a less intense precursor next. Two problems are immediately evident: mass spectrometers usually select the most intense precursor for fragmentation first. During the fragmentation cycle, the machine is blind to all other ions that elute and chances are real that a less intense peak has completely eluted from the column by the time fragmentation of the first precursor completes. Secondly, very abundant peptides can elute over a broad time interval, which can be several times larger than the period of blindness that was set on the mass spectrometer. As such, multiple fragmentation spectra will still be recorded for this peptide while ignoring less intense, co-eluting ions. With static exclusion a run is fully completed (often using dynamic exclusion within that run) and subsequently the full list of m/z values for identified peptides is presented to the

mass spectrometer. The run will then be repeated and the machine will ignore all precursors in the exclusion list. Again, two problems can be found. First of all, when the mass spectrometer cannot deduce charge state correctly, the exclusion list needs to contain at least 2 or 3 different m/z values for each precursor. This leads to an increase in the number of possible matching peptides, including peptides that have not been identified. These novel peptides will thus be erroneously ignored and cannot be detected. Second, the analysis time is doubled using this approach while the sample needs to be halved.
Finally, the exclusion list approach cannot alleviate the problems of quenching that occur for MALDI sources [Gevaert *submitted*].

- Selection of a subset of peptides from the digest prior to separation [Gygi 1999, Spahr 2000, Wang & Regnier 2001, Oda 2001]
  By selecting a subset of the total number of peptides from the initial proteolytic mixture, the complexity can be greatly reduced. This reduction can also be tailored through the use of an appropriate selection criterion (see table 1). The efficiency of the separation step(s) is enhanced by performing them on a less complex mixture, requiring less steps and thus minimizing sample-handling losses. Since there are less peptides presented to the mass spectrometer over a given time interval or within a certain spot on a MALDI target, higher recovery rates can be expected (a higher percentage of peptides presented are analysable in detail by the mass spectrometer). This increased sensitivity will subsequently allow better sample coverage, especially for low abundant proteins or hydrophobic proteins. An important problem with selecting peptide subsets is that the increased sample coverage advantage can be lost when peptide fragmentation spectra cannot be assigned to database entries. Indeed, selecting peptides from the mixture implicitly reduces the information redundancy at the peptide level. This is illustrated in figure 20 for various peptide selection techniques. It is therefore vital for these techniques to make every analyzed peptide count.

a)



b)

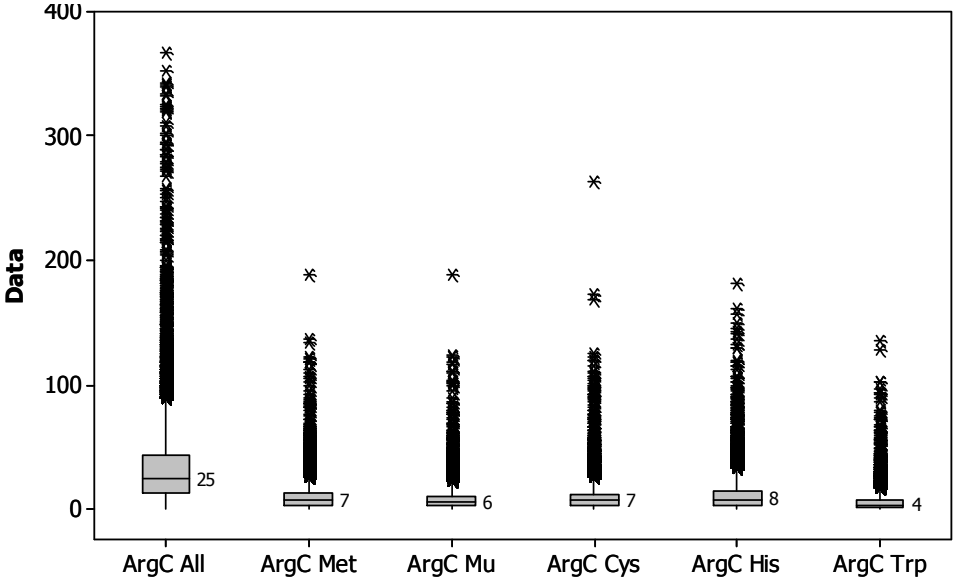


Figure 20: Boxplots of the number of in silico predicted retrievable peptides per protein using different sequence-based peptide selection strategies in SWISS-PROT (dd. 24 January 2006) for (a) trypsin and (b) Arg-C digests. The median for each boxplot is indicated. 'Mu' denotes methionines without taking initiator methionine into account.

1.4.3. The issue of undersampling

From the previous section it is clear that peptide selection has the benefit of dramatically reducing complexity while maintaining high theoretical sample coverage. It has also been established that this sample coverage is less robust because of lower redundancy at the peptide level (lower protein sequence coverage) requiring higher identification efficiency in order to obtain sufficient sample coverage. There is another issue with regards to sample coverage and peptide selection that remains to be discussed, however.

Whenever a sample is analysed by any proteomics technique, it will not be completely covered in one single analysis. This problem can be described as *undersampling*: the analysis method samples only a fraction of the complete sample for detailed analysis and thus identification. One of the problems of 2D-PAGE based methods as highlighted in section 1.1.4 above for instance, is that these approaches consistently fail to pick up certain classes of proteins. Sample coverage thus reaches a nearly unbreachable threshold, regardless of the number of times the analysis is applied.

Peptide-centric approaches (peptide-subset isolating as well as non-isolating techniques) provide a different case, however. It has been reported that with each new iteration, repeated applications of an analysis technique reveal novel proteins, together with redundant identifications of proteins identified in one of the previous iterations [Liu 2004].

In order to show how non-selective and selective peptide-centric strategies fare over multiple iterations with regards to coverage, a simplistic but revealing thought experiment is conceived: consider a complete digest of a sample consisting of 100.000 peptides. Now suppose that the LC-coupled ESI mass spectrometer used for producing fragmentation spectra has an upper limit of producing 8.000 peak lists over the gradient applied on the LC column. Also suppose that these peak lists are spread randomly across all peptides present in the mixture. This implies that in optimal conditions we only cover 8% of the peptides in the sample in a single iteration. Now consider a second analysis attempt, employing the exact same set-up. We now have 8.000 ‘known’ peptides and 92.000 ‘unknown’ peptides in the mixture. Since our mass spectrometer samples in an unbiased way, it can be expected that, of the 8.000 fragmentation spectra recorded, 7.360 (94%) will be derived from the 92.000 unknown peptides, and 640 (6%) from the 8.000 known peptides. This leads to a total number of 15.360 ‘known’ peptides, reducing the number of unknown peptides to 84.640 and reaching sample coverage of a little over 15%. Applying successive iterations will yield an ever dwindling number of novel identifications as the ‘unknown’ fraction becomes smaller and the unbiased mass spectrometer is less prone to sample from among them.

Of course, the limited capacity of our mass spectrometer is something we cannot change, but we can reduce the original sample complexity by selecting for a subset of peptides. If, for instance, we select for cysteine or methionine-containing peptides, we reduce complexity by roughly 75% (see table 1 and the boxplots in figure 20 above). This yields an initial sample complexity of 25.000 peptides. The 8.000 peak lists recorded by the mass spectrometer now not only provide a larger sample coverage in the first iteration, but also keep the coverage increase from dropping too quickly in subsequent iterations.

The N-terminal COFRADIC approach goes one step further and theoretically reduces each protein to a single N-terminal peptide, which would decrease sample complexity

even further to 3% of the original (SWISS-PROT has an average of 34 tryptic peptides per protein; see section 1.1.5). In practice, pyro-glutamate and proline-starting peptides are also picked up, so we will (somewhat pessimistically) state that we have 10.000 peptides in the mixture. It is clear that the sampling efficiency of our mass spectrometry (with its limited capacity) is now 80%!
The theoretical results of this thought experiment are visualized in figure 21.

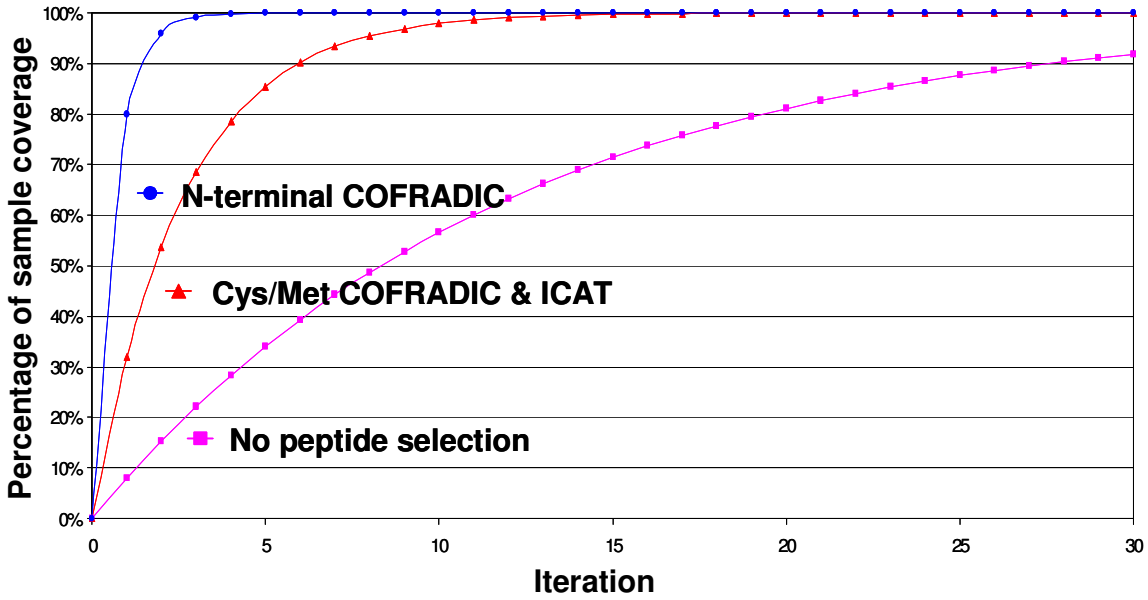


Figure 21: Sample coverage over successive iterations for peptide selecting and non-selecting techniques.

Of course, the above thought experiment lacks certain key features of real-life experiments. Peptide copy number is not taken into account nor the difficulty in eluting certain peptides on an RP-column or sequence-specific variability in the ability to generate good fragmentation spectra, to name but a few. Although these additional parameters would lower the absolute performances, it is also clear that they would affect the three strategies discussed above equally and therefore would probably not significantly alter their relative performance. The thought experiment thus provides a useful way to clearly show that, because of the effect of undersampling, a significant reduction of sample complexity is a very desirable characteristic of any peptide-centric technique that aims to achieve decent sample coverage over a minimum number of iterations.

1.5. Informatics challenges for a high-throughput proteomics laboratory

Data generation in a high-throughput proteomics lab very quickly reaches a volume that can no longer be supported by manual management of spectra and identifications. This is illustrated in figure 22, where the total number of fragmentation spectra in the ms_lims database of the proteomics laboratory in Ghent is shown for each month over a two-and-a-half year interval. Note that this growth curve corresponds to an average increase of approximately 38.000 spectra per month. Due to this high volume of data generated, a high-throughput proteomics lab today is confronted with at least three distinct software issues (and corresponding user roles): data integration, data processing and data analysis. Ideally, the results of the research will be published, raising a fourth software issue: data dissemination. We will discuss each of these separately in the following four subsections.

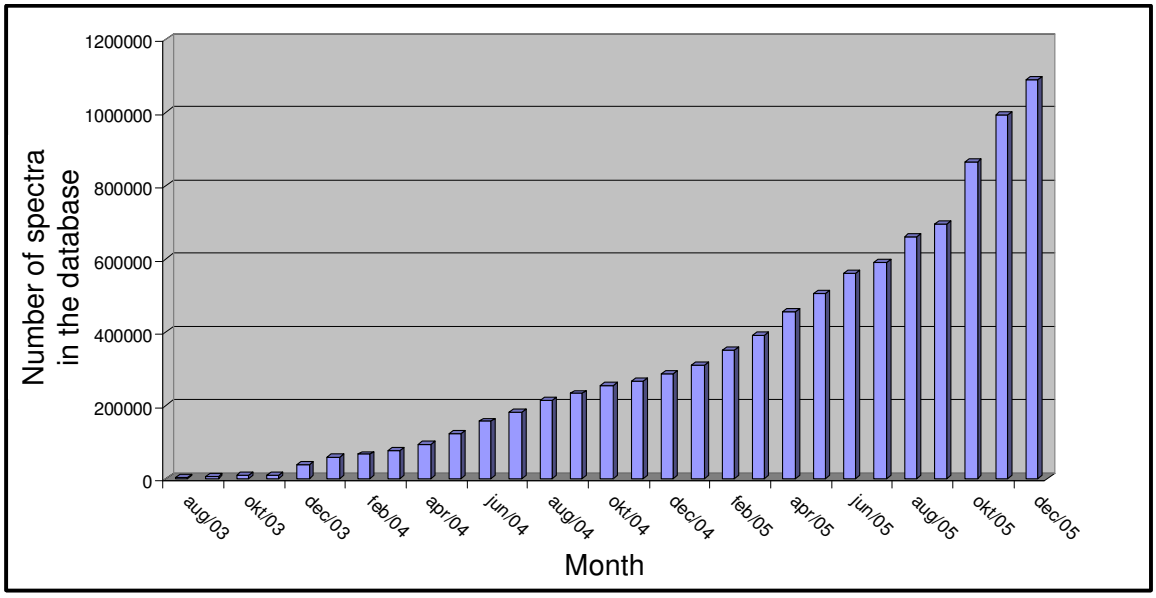


Figure 22: Number of spectra in the ms_lims database of the proteomics group in Ghent over time.

1.5.1. Data integration

It is clear that the flood of proteomics data necessitates the transition from manual management of the data produced to the use of specialized software for automated handling and structured storage. The software in question must first and foremost allow transparent and straightforward integration of the data presented by different producers, ideally converting this data to a common format in a shared infrastructure.
The different primary data producers in the proteomics laboratory in Ghent are outlined in figure 23. This set-up can be considered typical for many proteomics labs, as it is often the case that various instruments of various manufacturers are combined on the basis of the complementarity of their respective strong points.

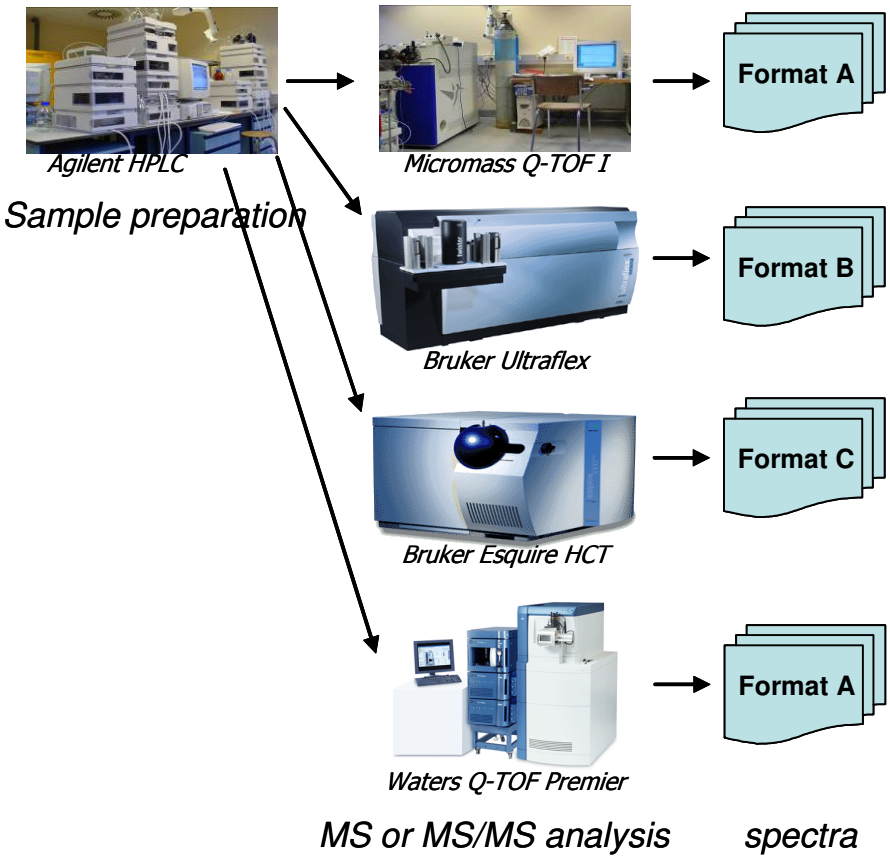


Figure 23: Schematic representation of the data producers in the proteomics lab in Ghent (situation dd.30 January 2006).

It is clear from even this rough schema that the different producers nearly all produce different output formats (this is even the case for machines from the same vendor!). Additionally, each machine has a different instrument management software and a different operator yielding different file management strategies for each machine as well. Software that successfully integrates the data produced by these machines therefore has to be tuned to each specific instrument while providing all operators with a consistent interface. The output of this software should be provided in a common format and stored in a common structure so that downstream applications need not concern themselves with the disparities in original formatting and structure. Finally, the integration software must also be easily extensible; whenever a new instrument is added to the pool of data producers, the amount of time and effort to include this new data source should be minimal.

1.5.2. Data processing

Data processing can encompass a whole range of possible manipulations of the data. The most essential of these is the step from spectrum to identified peptide and further to identified protein. This task is usually delegated to commercially or publicly available software such as SEQUEST [Eng 1994], Mascot [Perkins 1999] or X!Tandem [Fenyo 2003]. The challenge for the data management software here lies in integrating the

submission of spectra and retrieval of identifications from the third-party software. Additionally, some post-processing is usually required to filter out identifications that are deemed incorrect and to assign protein identifiers to reported peptides in a more apt way. These latter tasks can be incorporated into the data management software, or can be derived from third-party applications such as PeptideProphet [Keller 2002], DTASelect [Tabb 2002], CHOMPER [Eddes 2002] or ProteinProphet [Nesvizhskii 2003].

1.5.3. Data analysis

When the source data has been converted and stored, subsequently submitted to a third-party identification algorithm and the results have been retrieved, parsed and stored, the now complete dataset needs to be presented to the operator or the scientist in order for them to extract value from the data. As there are many possible ways to analyse the data, there can be no single data analysis application that can take care of handling all possible questions a data consumer might have. For this reason, the data analysis software is usually composed of a suite of several (independent) modules, each of which retrieves and analyses those sections of the total data storage it deems relevant. These modules are the most variable part of the entire software package. Apart from a few basic data retrieval and visualisation applications, most of the modules will have been tailored to answer a specific question at a specific time. When the relevance of this question has diminished, the specific application may fall into disuse. Of course, at any specific time in the future, the question might gain in importance again, prompting a data consumer to revert to using the application once more. The relevance of this is that an analysis application once programmed and released in production, will usually require maintenance even if it seems to have fallen into disuse.

1.5.4. Data dissemination

Data dissemination is the Achilles heel of proteomics research [Prince 2004, Rohlff 2004]. This tendency to hold back or obfuscate data used in published results is particularly sad as any scientific undertaking since the Enlightenment hinges on the free communication of results. Indeed, it can be stated that data is the blood of the scientific community and that papers are the pumps that should drive its circulation.
There are numerous reasons why so little data has been made available and so little of the available data is readily accessible. These reasons fall into two broad categories: logistical and sociological (for lack of a better term) reasons. The logistical reasons are outlined and discussed next. It is of course beyond the scope of this dissertation to thoroughly analyse the sociological phenomena that occur in the proteomics scientific community and these will not be discussed here.
In order to publish a large amount of data (e.g. tens of thousands of peak lists, yielding thousands of peptide identifications which in turn support several hundred protein identifications) in a queryable, human and machine readable way poses three distinct challenges.
The first problem centers on *local data management*. If data is managed manually from start to end, it is usually impossible to back-trace from protein to supporting peptides to original peak lists. Even if data has been collected automatically using third-party

software, it is often impossible to export the required data in a meaningful format for general dissemination because the software does not contain this feature by default and its data is insufficiently accessible to provide a solution in-house.
The second problem concerns *standardization*. Standardization works on two separate levels. The first part of any standard describes the minimal amount of information that needs to be reported. This is in fact the most difficult issue in the entire data dissemination discussion. A young and quickly changing field such as high-throughput proteomics makes it extremely dangerous to attempt to set certain rules in stone. The use of specialized protocols and very different data processing pipelines only adds to this problem. In fact, having every single protocol and data processing step documented does not necessarily result in allowing reproduction of the analyses performed in order to validate the results, as specific steps may not be available to other researchers[12].
The second part of any standard concerns data formats. Once the specific data that should be communicated has been decided upon, a format needs to be found that can accommodate this information in a universally readable way.
The Human Proteome Organisation (HUPO) [Hanash 2002] has created the Proteomics Standards Initiative (PSI) [Orchard 2005] to tackle both standardization problems. Minimal reporting guidelines are being drafted in the Minimal Information About a Proteomics Experiment (MIAPE) documentation and standardized formats for mass spectrometry data (mzData) and data analysis (analysisXML) are under active development. Other published mass spectrometry data standards are also available [McDonald 2004, Pedrioli 2004]. Additionally, certain journals have also published guidelines aimed at setting standards [Carr 2003, Bradshaw 2005, Wilkins 2006] that submitted manuscripts should minimally adhere to. Of course, this heterogeneity of standards bodies (HUPO PSI, several journals) will need to be resolved in full before real progress can be made in this matter.
The third problem is the availability of a centralized, *publicly available repository* for the published data. Solutions relying on private websites of researchers or laboratories suffer from three problems: (1) they are notoriously unstable over time and are quite often discouraged by journals for this reason[13], (2) they can never provide a uniform interface or data model to potential users and (3) they delegate the cost of site construction, site maintenance and bandwidth to the authors, discriminating against the less affluent and exacerbating the already short website half-life discussed in (1).
The only real solution is to provide steadily funded, centralized data repositories at specialized institutes that can collect, publish and maintain data submitted by authors and present users with a uniform interface. This approach has of course been in place for years for nucleic acid and protein sequence databases and has repeatedly proven to greatly aid researchers worldwide.

[12] Examples of these are prohibitively expensive reagents, the use of a custom-built apparatus at one or more steps and the use of proprietary data processing algorithms to achieve certain results.
[13] An example of a similar distrust in private website stability can be found in the 'Instructions to Authors' of the BioMedCentral journal BMC Bioinformatics, where authors are encouraged to submit the software applications they describe in their manuscripts as supplementary information to the journal rather than distributing the software through their own homepages.

1.6. The choice of a programming language

"It's not difficult to look at computer languages and see which ones are trying to be modern by driving something into the ground. Think about Lisp, and parentheses. Think about Forth, and stack code. Think about Prolog, and backtracking. Think about Smalltalk, and objects. (Or if you don't want to think about Smalltalk, think about Java, and objects.)"

- Larry Wall

Whenever any software development is to take place, one of the first questions raised concerns the programming language to use. This choice is often disregarded as trivial (one simply picks the language one is most proficient at using) yet carries far-reaching consequences. When considering the large software systems which are designed to tackle the points raised throughout section 1.4, it becomes clear that the initial choice of the programming language will need to be carried through for many years, possibly suffering changes in programmer staff over that time period. The selection of a suitable programming language then becomes considerably less trivial.

Any software that is put into production needs to be fixed, maintained and updated. Each of these aspects requires the modification or addition of source code. It is therefore of the utmost importance that key software systems (such as a data management and analysis system) are written in clear, understandable and documented code.

Java was the language of choice for the software developed in the course of the work outlined in this dissertation. Java has the benefit of being a very robust object-oriented language with the highly interesting property of providing cross-platform binaries. This means that the compiled code itself can be distributed and run on widely diverse platforms and architectures without any problems and with only minor (if any) differences in the presented graphical user interface. It is clear that in any academic setting this is an advantage, especially when one aims at sharing the software in question across the community. Additionally, Java contains powerful tools that allow in-code documentation to be compiled into external HTML-based documentation using a standardized format. Java also has an excellent track record concerning the completely transparent integration with enterprise-level tools such as databases. In practice this means that a system built to run on one relational database platform can be run without any code-modification on a completely different relational database engine. Finally, the popularity of Java and the elegance of its object-oriented design have made it a favourite in undergraduate programming courses. As such, there are many Java-knowledgeable programmers available, lessening the sting of the infamous '*hit-by-a-bus-factor*'[14].

[14] This factor is a somewhat whimsical way in which software developers refer to 'how many people can get hit by a bus before the project gets in trouble'. Dark humor aside, the statement often rings true: the situation where a single developer leaving a team (for whatever reason) wrecks havoc on the whole system is not an uncommon one.

1.7. Relational databases

1.7.1. Introduction

Relational databases have been the mainstay of structured data storage for decades. They offer a very fast, reliable mechanism for inserting, updating and retrieving data. The ability to query these databases through an official standard language (Structured Query Language, or SQL) only adds to their usefulness. Enterprise-level relational database engines are available from many vendors (notably Oracle, IBM (DB2), Microsoft (SQL-server) and Sybase), but many free and/or open source solutions are available as well (notably MySQL and PostgreSQL). These latter solutions have become feature-rich and highly competitive systems in their own right, bringing the complete power of a relational database system into the reach of everyone.
Relational databases have a drawback however: object-oriented models are not readily compatible with a relational schema. Key object-orientation concepts such as inheritance and typing are not straightforward to implement in a relational schema. As an alternative, object-oriented databases have been created, but these tend to be slower and far less manageable than relational stores. They also lack a uniform query language and suffer from far less support in popular programming languages.
For the majority of data-centric applications however, relational data storage is sufficiently expressive to allow complete modeling of the data structure without having to resort to arcane manipulations.
Relational databases are constructed from tables. These define a number of columns, each of which carries a certain type of data. Whenever data is added to such a table, rows are inserted. There are two special columns which are instrumental to creating relations between otherwise unconnected columns. These are the 'primary key' column(s) and 'foreign key' column(s). Whenever more than one column is used to create such a key it is called a 'compound key'. For the sake of simplicity compound keys will not be discussed further here. A foreign key is a column that references a column in another table, typically the primary key for that table. It is this referencing that creates the connection between the tables. Whenever a reference between keys is followed, it is called a 'join' of the corresponding tables.
In modeling a relational database schema one can make use of three basic relations, all defined by primary key-foreign key referencing. These atomic elements of relational database design are discussed in the following subsections.

1.7.2. One-to-one mapping (1:1)

A one-to-one mapping (often abbreviated to '1:1') consists of two tables that are interrelated in such a way that a single row in one table is matched by a single row in the other table. This relationship is reciprocal. It is rarely ever used as one-to-one relationships can be readily collapsed into a single table. They sometimes come in handy to overcome potential limitations of the underlying database engine[15]. An example for two simple one-to-one tables is depicted in figure 24. Note that they share their primary

[15] A common example of this is the limit on the number of columns that a table can accommodate.

key and that this will be the join column. This relation therefore does not require foreign keys.

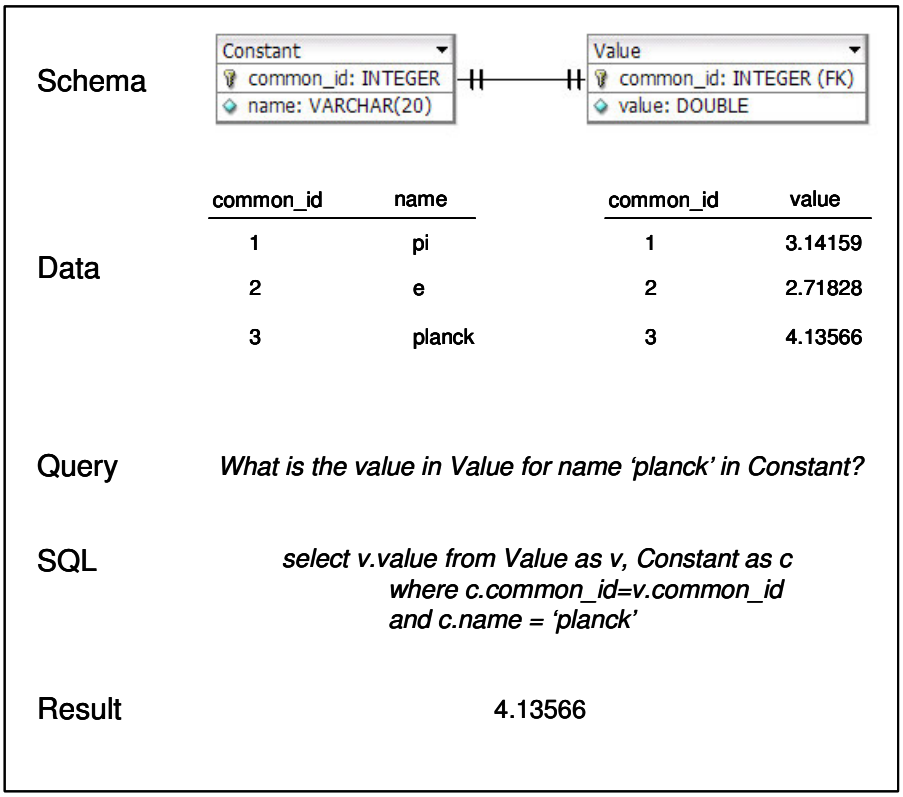


Figure 24: Example for a 1:1 relationship between tables.

1.7.3. One-to-many mapping (1:n)

A one-to-many mapping is often abbreviated to '1:n'. This relation can safely be called the workhorse of most relational database schemas. Indeed, it will be argued in the next section that many-to-many mappings are actually a peculiar combination of two one-to-many mappings.

One-to-many mappings rely on foreign keys to work their magic. The primary key of a single row in table 'A' can be referenced by the foreign keys of many rows in table 'B', allowing a data consumer to retrieve all matching records ('many') in table 'B' for a specific row in table 'A' ('one'). When applied in the inverse direction, a single row in table 'B' will only link a single row in table 'A'. This relationship is thus clearly unidirectional. An example for two simple tables is given in figure 25.

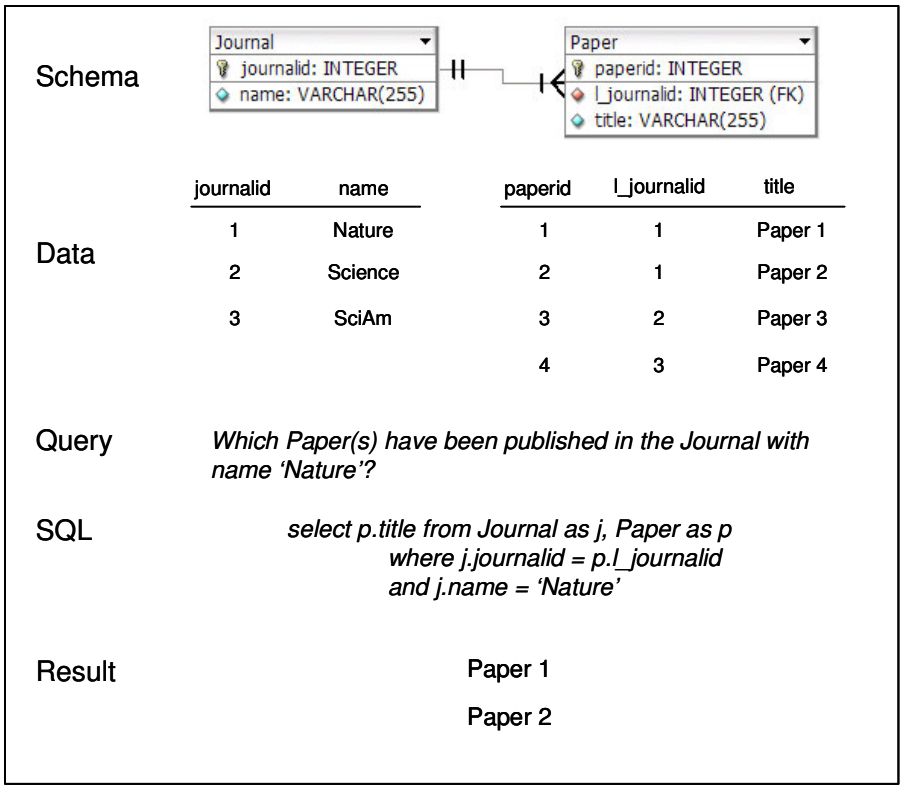


Figure 25: Example for a 1:n relationship between tables.

1.7.4. Many-to-many mapping (n:m)

Many-to-many mappings, or 'n:m' in short are the third basic element of relational data storage. In this construct a single row from table 'A' can be joined to many rows in table 'B', but it is also possible to link many rows in table 'A' from a single row in 'B'. The actual relations retrieved need not be reciprocal however. The key to this element is an extra table often referred to as the 'indirection table'. This table consists of only two foreign key columns in its most simplistic form. For two tables 'A' and 'B' to have a many-to-many relationship, a third table 'C' is thus required. Table 'C' will have at least two columns, one of which will contain a foreign key referencing table 'A' (let us call it 'c1'), whereas the other will be a foreign key referencing table 'B' ('c2'). It now becomes possible to start with the primary key of a single row in table 'A', find all rows in table 'C' that have a foreign key in 'c1' referencing this primary key and then use the foreign keys in 'c2' for these rows in table 'C' to locate the corresponding rows in 'B'. Note that this boils down to applying a one-to-many relationship and an inverted one-to-many relationship in sequence. Also note that the inverse operation (going from a single row in table 'B' to all matching rows in table 'A' via the foreign keys of table 'C') will also work. An example for two tables and a single, minimalist indirection table is given in figure 26.

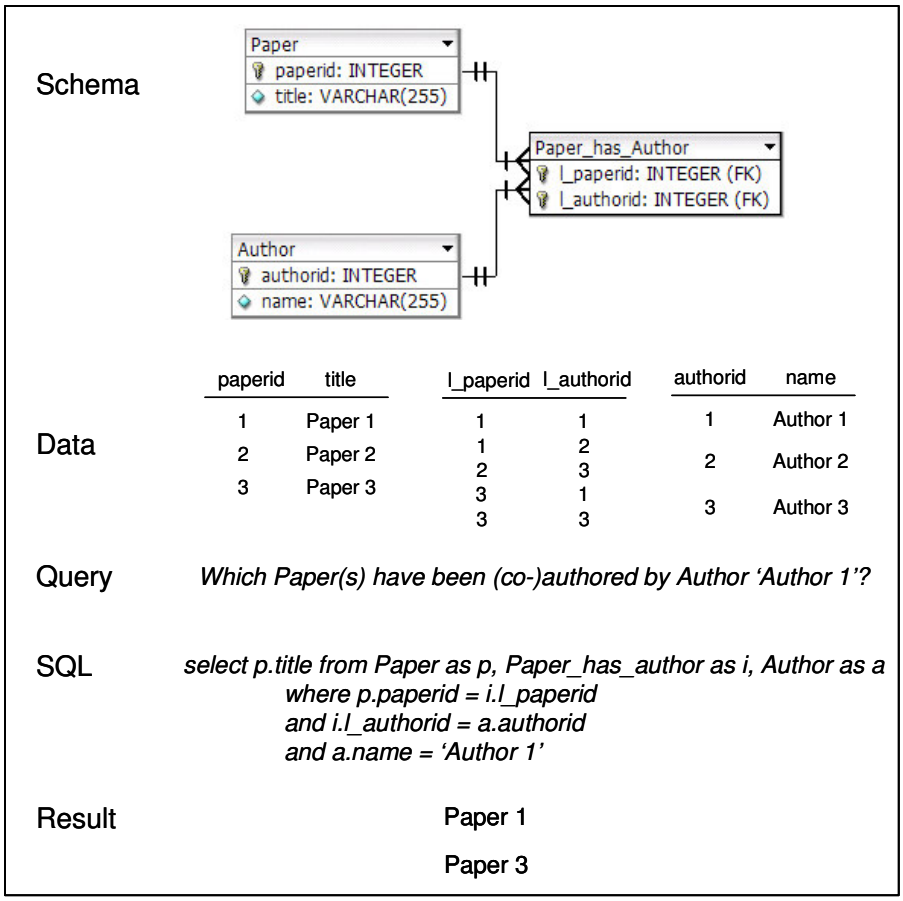


Figure 26: Example for an n:m relationship between tables.

2. Results

2.1. DBToolkit: creating the right database for the job

"The solutions all are simple... after you have arrived at them. But they're simple only when you know already what they are."

- Robert M. Pirsig, *Zen and the Art of Motorcycle Maintenance*

Peptide identification from fragmentation spectra relies on three main elements: the quality of the MS/MS spectrum, the quality of the search engine and the quality of the search base against which the spectrum is matched.
One of the most crucial elements is the search base used for peptide identification. Indeed, if the sequence of the peptide that led to the recorded fragmentation spectrum is not contained in the search base, it will be impossible to find. This problem becomes particularly pressing for those peptide-centric approaches that specifically pick up processed peptides such as N-terminal COFRADIC as the site of processing will only sporadically conform to a standard enzymatic site. An illustration of this problem for (partial) cleavage of a caspase substrate can be found in figure 27. It is immediately clear from this representation that, when the blue peptide is missed, the entire cleavage event is missed. This nicely illustrates the principle put forward in section 2.1: we must make every peptide count.

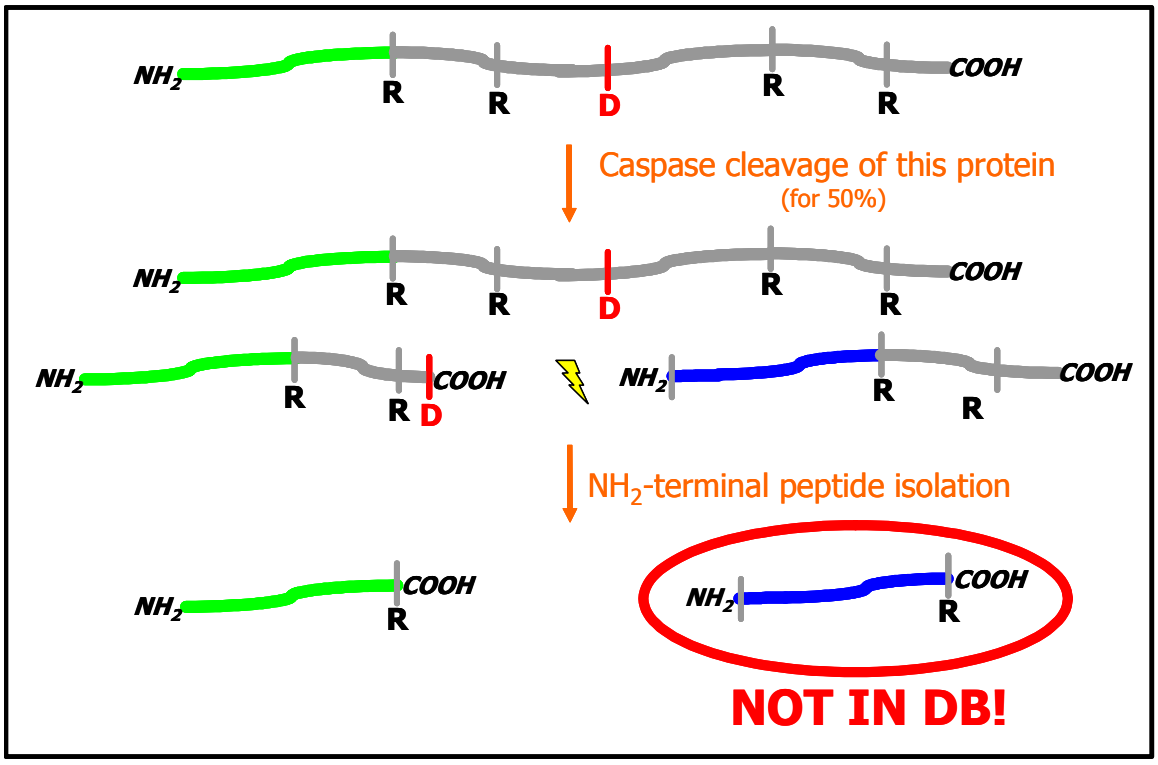


Figure 27: Partial cleavage of a caspase substrate, resulting in a novel N-terminal peptide that is not present in regular Arg-C digested databases.

In this section we will introduce the DBToolkit software application, which was designed to aid in making every peptide count. We first discuss the processing of a protein sequence database in terms of information content and information ratio. We will then proceed to illustrate the usefulness of corrupting the information content of a database and we will highlight the part of the design of DBToolkit that gives it the essential ability to be easily adapted to changing needs by third-party developers. The section is completed by inclusion of the published DBToolkit paper and illustrations of its usefulness in published applications of peptide-centric databases.

2.1.1. Transforming protein sequence databases into peptide databases

In shifting the focus from gel-based to gel-free methods, the unit of analysis shifted from the protein to the peptide (hence the term 'peptide-centric' proteomics). The sequence databases used as search bases for the identification algorithms however, necessarily remain protein-centric. DBToolkit was designed to solve this discrepancy by allowing the processing of protein sequence databases into peptide sequence databases. Additionally, DBToolkit can enhance both the information content as well the information ratio of the peptide sequence database[16]. Strategies for the former will be discussed in section 2.2.1.1, the mechanism for the latter is detailed in section 2.2.1.2.

2.1.1.1. Enhancing the information content of a peptide database

We have illustrated above that the most interesting N-terminal peptides are often obtained after some form of *in vivo* processing, typically through a proteolytic event. We have also pointed out that a standard *in silico* protein database digest using trypsin or endopeptidase Arg-C will generally not generate these peptides and that a search algorithm such as Mascot will therefore not be directly able to consider these as potential matches. The end result is that the biologically most interesting peptides will be completely missed. There are some strategies available that attempt to correct this problem. First of these is the semi-specific setting for the *in-silico* proteolytic enzyme which requires only one terminus (either the C- or N-terminus) to correspond to a correct enzymatic cleavage. Thus, the following sequence:

>Protein 1
NARTMA

will yield the following list of peptides by applying semi-specific tryptic cleavage (note the subscript 'c', indicating a correct tryptic terminus):

$_cNAR_c$
$_cTMA_c$
AR_c
R_c
$_cNA$

[16] Interestingly, one of the other highly useful features of DBToolkit is that it can also corrupt the information content of a sequence database, which is explained in section 2.2.2.

$_cN$
MA_c
A_c
$_cTM$
$_cT$

A more brute-force approach is the Mascot 'No Enzyme' setting. This setting effectively considers all possible peptide sequences for a given polypeptide. For our example protein above, we would get the following non-redundant peptide list:

NARTM
NART
NAR
NA
N
ARTMA
RTMA
TMA
MA
A
ARTM
RTM
TM
M
ART
RT
T
AR
A
R

It is clear that the semi-specific approach will most probably cover every possible form of processing and that the No Enzyme setting will cover every possible sequence that can be derived from the parent protein. The information content of the database (which we will consider to be the number of unique sequences in that database) is thus increased by these settings.
This increase in information comes with a price, however. Mascot is a probabilistic algorithm and relies on the number of amino acids in a sequence database to estimate its information content (see section 1.2.1.2). When applying semi-specific or No Enzyme cleavage, the corresponding increase in information content must be taken into account when estimating the chance of random matching. We can easily plot the effect on a simple chart for the human subsets of SWISS-PROT, IPI and the NCBI non-redundant database (figure 28).

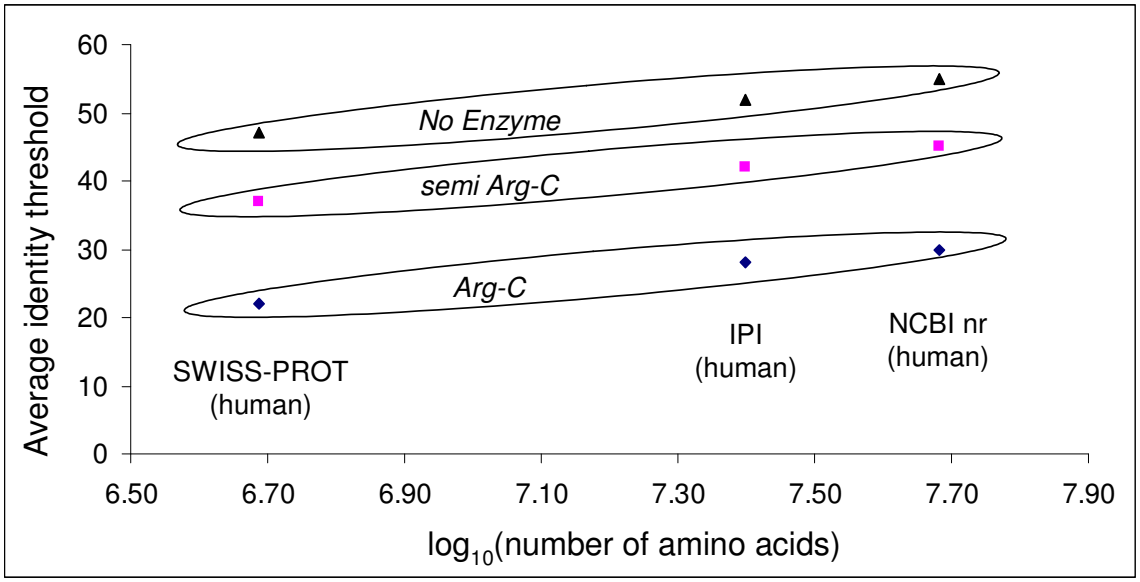


Figure 28: Increase in identity threshold when applying semi-specific enzyme or No Enzyme setting in Mascot

It is obvious from this figure that the resulting penalties are steep. It is therefore to be expected that, even though the information content of the databases can be raised to an adequate level using these settings, the chances of correctly linking a peptide sequence to its fragmentation spectrum are somewhat reduced due to the high identity threshold score. It is also interesting to note that the search space employed is not necessarily optimized with regards to what can be reasonably expected from the protocols used. Let us first consider the example of regular N-terminal COFRADIC. As this technique allows the efficient selection and retrieval of the N-terminal peptide of a protein, it is a very useful tool to study *in vivo* protein processing. The peptides resulting from this protocol will conform to Arg-C cleavage on their C-terminus (thus ending in arginine) as described in section 1.1.8.3. The N-terminus of part of the peptides will conform to Arg-C cleavage as well, but another part (those that correspond to the N-terminus of a protein that was processed during maturation) of the peptides in the mixture will carry an N-terminus that results from a completely different proteolysis event. We can therefore make the *a priori* prediction that the N-termini can not be clearly specified, but the C-termini of the peptides should end in arginine. Now let us reconsider the peptide list created by an *in silico* semi-specific Arg-C cleavage as given above, only now the peptides which we would accept as reasonable are given in bold face:

$_c$NAR$_c$
$_c$TMA$_c$
AR$_c$
R$_c$
$_c$NA
$_c$N
MA$_c$
A$_c$

$_cTM$
$_cT$

It is clear that, of the ten possible peptide sequences, only six can be considered plausible. Another way to say this is that the truly useful information content of this database is only 60% of the total information content. Since we have established that the identity threshold of a database relies on its total information content, it is clear that the identity threshold will be overly pessimistic for our purposes.
It is actually straightforward to provide a solution for this problem. When we apply our *a priori* knowledge to the *in silico* protocol, we can create a peptide sequence database that contains all regular Arg-C peptides next to possible N-terminally truncated versions of these. To continue our above example, we would create the following peptide list:

$_cNAR_c$
AR_c
R_c
$_cTMA_c$
MA_c
A_c

Note that this list is simply a rearranged version of the bold entries in the list above. Applying this algorithm, optionally complemented with mass limits, will then produce a more informative peptide-centric database, which can be used as an optimized search base for regular N-terminal COFRADIC (figure 29).

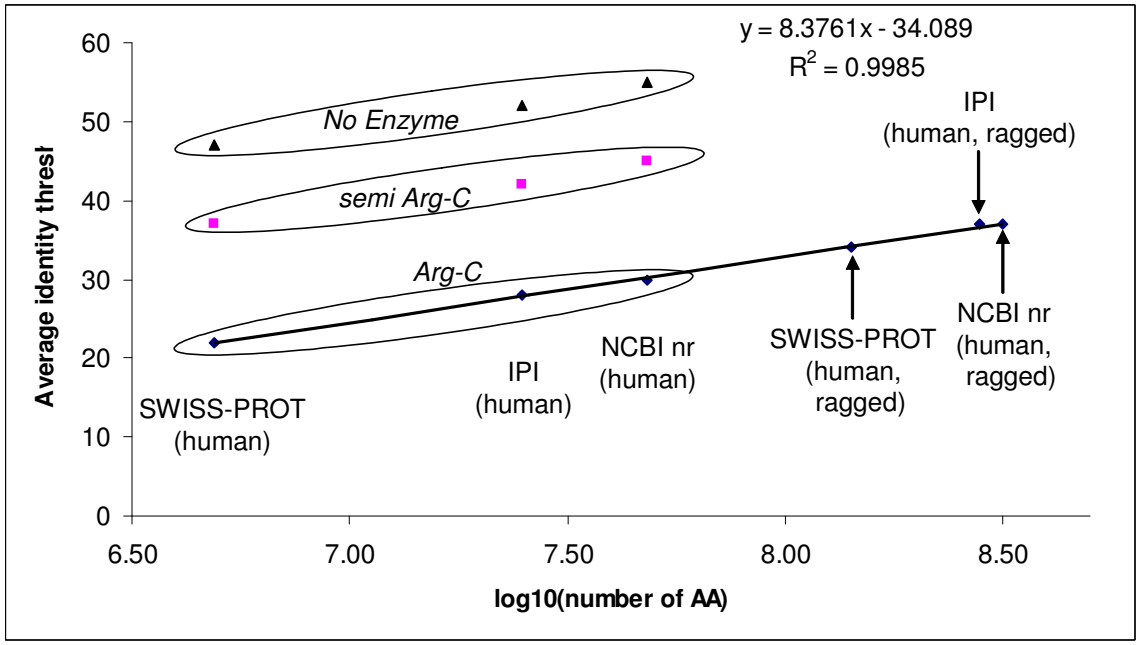


Figure 29: threshold scores and sizes compared for regular enzymatic databases, semi-specific and No Enzyme databases and non-redundant ragged databases output by DBToolkit

A schematic representation of the actual algorithm can be found in figure 30.

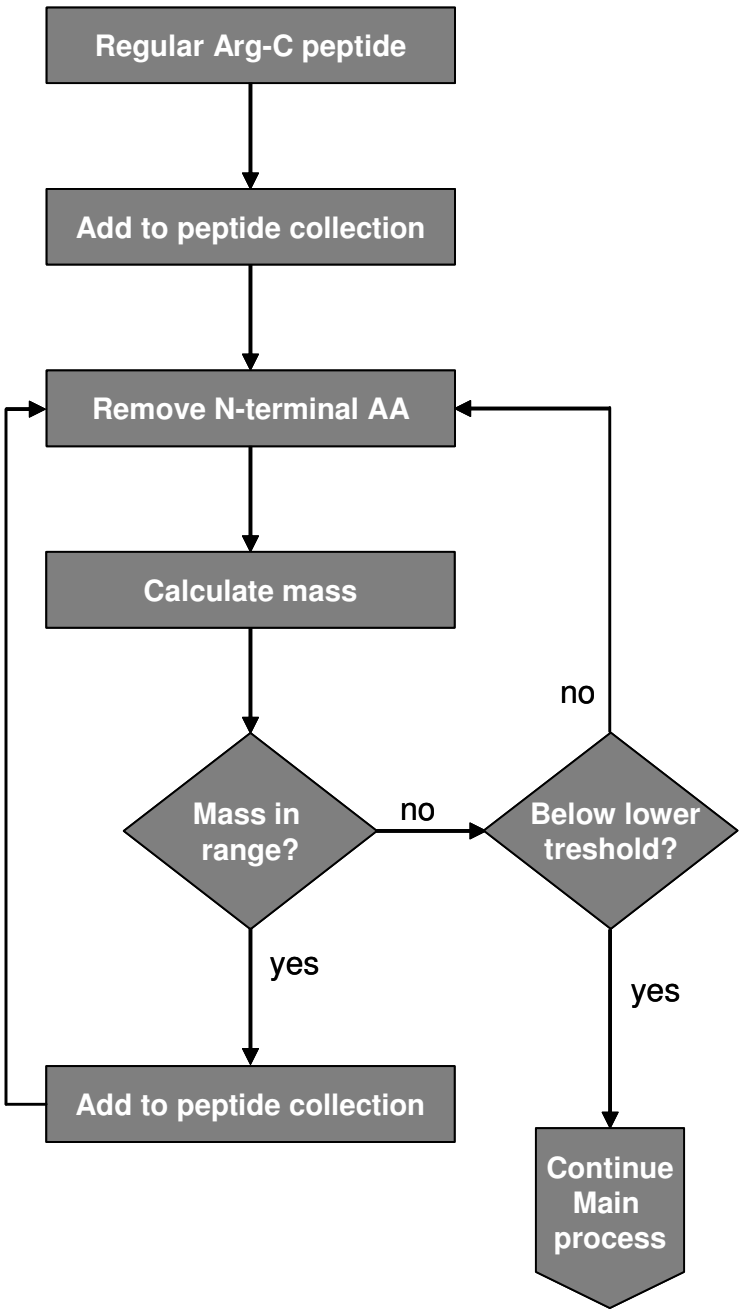


Figure 30: N-terminal 'ragging' of Arg-C peptides. The 'Main process' cycles all protein entries in the database and generates regular Arg-C peptides that are fed to this ragging process.

2.1.1.2. Enhancing the information ratio of a peptide database

"Pessimists, we're told, look at a glass containing 50% air and 50% water and see it as half empty. Optimists, in contrast, see it as half full. Engineers, of course, understand the glass is twice as big as it needs to be."

- Robert Lewis

One of the problems encountered when transforming protein sequence databases into peptide-based sequence databases has to do with redundancy. Usually protein sequence databases are deemed non-redundant when each protein sequence in the database is unique. Proteins from large families with extensive sequence homology such as members of the globin family tend to produce many identical peptides after proteolytic digest. The redundancy at the peptide level is thus generally much higher than for the original protein sequence database. This redundancy becomes problematic when one looks at the workings of probabilistic protein identification algorithms such as Mascot. Indeed, a large amount of sequence redundancy at the peptide level implies that the total number of amino acids in the database becomes a progressively more optimistic estimate of the information content of that database as the redundancy increases. Consider for example the following two-protein database:

```
>Protein 1
LENNARTMARTENS
>Protein 2
LENNARTMARTENT
```

and the six-peptide database generated by an *in silico* digest using Arg-C:

```
>Protein 1 (1-6)
LENNAR
>Protein 1 (7-10)
TMAR
>Protein 1 (11-14)
TENS
>Protein 2 (1-6)
LENNAR
>Protein 2 (7-10)
TMAR
>Protein 2 (11-14)
TENT
```

Even though the sequence redundancy at the protein level is 0%, the redundancy at the peptide level is 33% (two out of six peptides are redundant). Correspondingly, the information content of the Arg-C peptide database is equal to four unique peptides. If we now define the information ratio of a database as its information content divided by the total number of sequences in that database, we find an information ratio of 67% in the

above example. The relationship of this measurement with the number of amino acids in a sequence database is readily established:

$$I_r = \frac{I_c}{S_s} = \frac{I_c \cdot < L_p >}{S_{aa}}$$

Where I_r equals the information ratio, I_c the information content, S_s the total number of sequences, S_{aa} the total number of amino acids and L_p the average peptide length in a database.
It is clear that an ideal database would have an information ratio of 100%; i.e. every sequence in that database is unique and therefore informative, allowing the most efficient estimate for random matching from the total number of amino acids in the database.
To raise the information ratio of a database it is therefore only necessary to remove sequence redundancy. In the above example of the six-peptide database, we get the following list of unique peptides:

LENNAR
TMAR
TENS
TENT

The redundancy in these peptide sequences is obviously 0%, and the information content remains at four unique peptides. With an information ratio of 100% we now have an ideal search base to present Mascot with.
A good illustration of the real-life impact of removing peptide-level redundancy is the information ratio of the NCBI nr database (dd. 28th of January 2006) after Arg-C cleavage (allowing 1 missed cleavage), which amounts to a meager 33%! With only one in every three peptide sequences actually contributing to the information content of the database, clearing the redundancy becomes vital to allowing a probabilistic search engine like Mascot to correctly estimate the chances for random matching. This effect is also depicted in figure 31 for the ragged human subsets of the NCBI nr (dd. 28 January 2006), SWISS-PROT (24 January 2006) and IPI (24 January 2006) databases, digested by Arg-C.
Note that the NCBI nr database contains nearly double the amount of ragged human peptide sequences after digestion with ArgC, yet adds only a marginal fraction of novel sequences when compared to the human subset of the IPI database. This is not particularly surprising when one considers the composition of these databases (see section 1.2.2) and clearly highlights the importance of the choice of sequence database in protein identification.

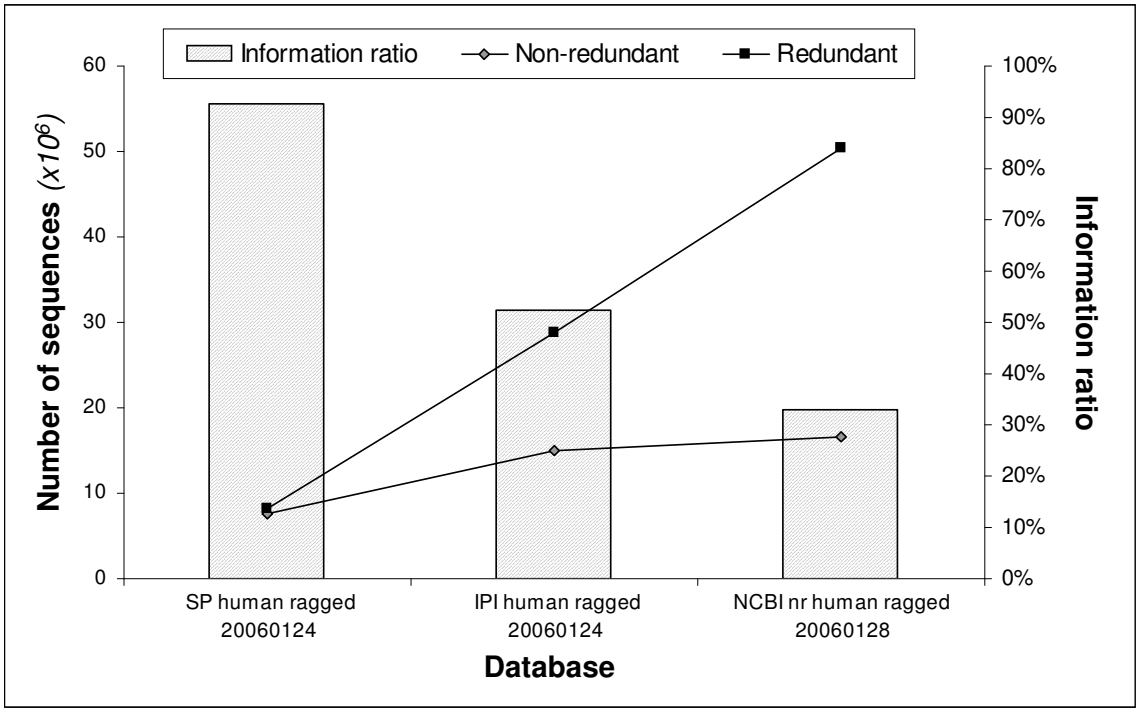


Figure 31: information ratio for ragged SWISS-PROT, IPI and NCBI nr human subsets. The line plots are matched to the left vertical axis, the bars to the right axis.

There is one important caveat regarding the removal of redundancy, however. The four sequences presented in the non-redundant list of the example above contain two *degenerate* peptides [Nesvizhskii 2005]. These peptides can not be clearly and unambiguously linked to a single protein sequence unless there is additional, conclusive evidence to perform an informed selection at the protein level[17]. This information is typically absent from peptide-centric approaches.We must therefore make sure we maintain all possible protein matches for these peptides in the non-redundant peptide database as none of these can be disregarded in favour of another. The final non-redundant peptide database will therefore be written by DBToolkit in the following way[18]:

```
>Protein 1 (1-6)^AProtein 2 (1-6)
LENNAR
>Protein 1 (7-10)^AProtein 2 (7-10)
TMAR
>Protein 1 (11-14)
TENS
>Protein 2 (11-14)
TENT
```

[17] Usually obtained from separation steps at the protein level. Examples are apparent protein MW and pI.
[18] Note that the '^A' notation is the standard notation to indicate multiple progenitor accession numbers in cases of sequence redundancy as employed by the NCBI.

The result is a peptide database that employs maximum parsimony at the sequence level while retaining full information on sequence origins in the header. Indeed, it can be said that we have performed a lossless compression of the original peptide database.

2.1.2. Corrupting the information content of a sequence database

"The importance of nonsense can hardly be overstated. The more clearly we experience something as 'nonsense', the more clearly we are experiencing the boundaries of our own self-imposed cognitive structures."

- Gary Zukav, *The Dancing Wu-Li Masters: An Overview of the New Physics*

The use of randomized, shuffled or reversed databases in establishing rates of false-positive identification for database search algorithms has been firmly established [Elias 2005, Stephan 2005]. We will discuss these three types of *decoy databases* and their properties next.

A reversed database is a sequence database in which each sequence has simply been replaced by the reverse of its sequence, e.g.:

```
>Protein 1
LENNARTMARTENS
>Protein 2
LENNARTMARTENT
```

becomes:

```
>Protein 1 (rev)
SNETRAMTRANNEL
>Protein 2 (rev)
TNETRAMTRANNEL
```

Note that this approach maintains both the amino acid composition of the original database[19] as well as the lengths of the peptides generated[20].

A second approach is the shuffled database, which transforms a database by randomly shuffling the amino acids per sequence. For our two-protein example database this could yield:

```
>Protein 1 (shuf)
TNTALEERNMSNRA
>Protein 2 (shuf)
NMERLANATERTTN
```

[19] The amino acid composition is also preserved at the level of the individual sequence

[20] Note that this statement is not entirely true. For instance, for trypsin, un-cleavable KP and RP reversed, gives cleavable PK and RP, thus changing the lengths of some peptides. For the grand majority of peptides however, the lengths are maintained.

Obviously, if the shuffling algorithm is well-designed, the actual resulting sequences will probably differ each time it is applied to a given sequence database. Note that, although this approach maintains amino acid composition, it does not maintain peptide lengths at the sequence level.
The third approach employs randomized databases. In this case, the amino acids constituting a sequence are replaced by randomly chosen amino acids. Usually, the algorithm can be instructed to maintain a certain amino acid composition or can substitute amino acids purely at random if desired[21]. It is clear that, although amino acid composition can potentially be maintained for the complete database, this is usually no longer the case for each sequence. Obviously, peptide lengths will not be maintained at the sequence level.
This third approach for generating randomized databases is not often employed. Usually shuffled and reversed databases yield a better decoy database as they maintain more characteristics of the original database. DBToolkit therefore only supports the generation of both shuffled and reversed databases.
From a purely conceptual point of view, the shuffled database maximizes the sequence variability of the decoy database when compared to the more conservative reversed database. The first benefit of shuffling is that *peptide families* (large collections of peptides with very homologous sequences) are nearly completely dispersed among all possible shuffled variants in shuffled databases (see figure 32). Secondly, palindromic peptide sequences are simply repeated in reversed databases, contributing to a higher degree of overlap between the source database and the reversed database.

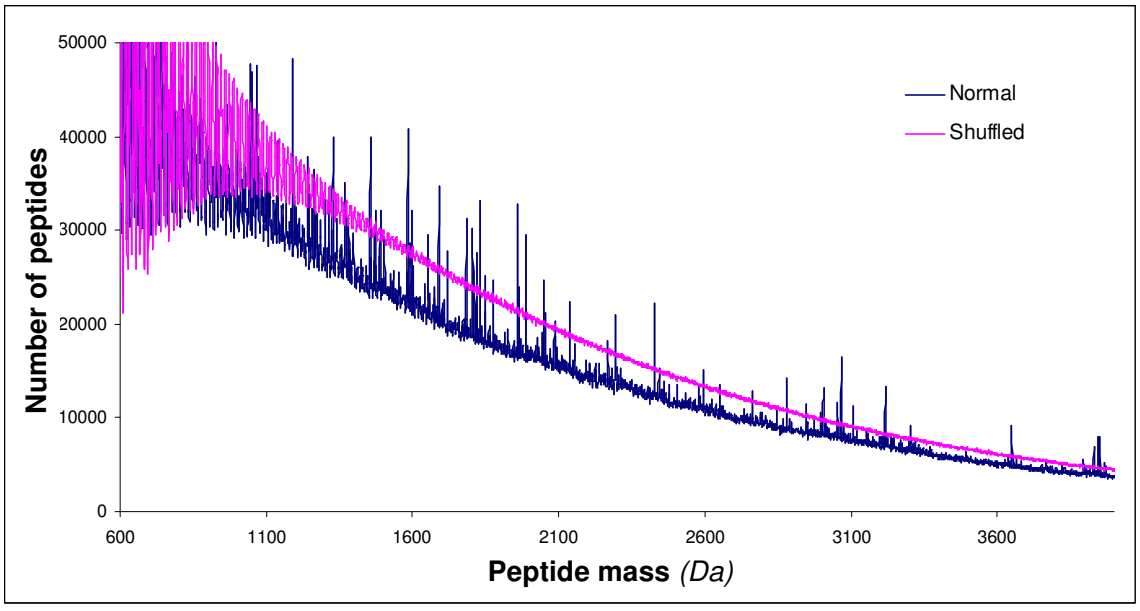


Figure 32: Mass distribution of all tryptic peptides (1 missed cleavage allowed) in the human subset of the NCBI nr database (dd. 27th of July 2003). Note the elimination of the various 'spikes' in the shuffled database.

[21] Purely random usually implies giving each amino acid a one-in-twenty chance of being picked to substitute at a given location.

The above points are also summarized in table 2.

Decoy mechanism	Composition maintained		Peptide lengths maintained		Potential overlap
	sequence	***database***	***sequence***	***database***	
Reversing	yes	yes	yes°	yes	low
Shuffling	yes	yes	no	yes	very low
Randomizing†	no	yes	no	yes	very low
*Randomizing**	no	no	no	no	very low

*Table 2: Comparisons of algorithms for the creation of decoy databases.° see footnote 19 above; † randomizing using a compositional bias; * randomizing using a one-in-twenty chance for each amino acid to be used as substitute.*

Mascot was tested using shuffled versions of the human proteins in SWISS-PROT and NCBI nr for two distinct datasets: the methionyl and cysteinyl COFRADIC peptides from human platelets (see section 2.5). The results are given in table 3.

Database	Spectra	Identifications		False (+)
		Regular	***Shuffled***	
SWISS-PROT	3565	1270	29	2.28%
	2665	586	29	4.95%
NCBI nr	3565	1137	28	2.46%
	2665	481	21	4.37%

Table 3: Estimation of false positives for Mascot at the 95% confidence interval using shuffled decoy versions of the human entries in the SWISS-PROT (dd. 25 July 2003) and NCBI nr (dd. 27 July 2003) databases for two distinct datasets.

It is clear that Mascot does not exceed its self-imposed maximum of 5% allowed false positives. Additionally, it seems that the quality of the dataset (as can be estimated by the identification ratio[22]) has more bearing on the false-positive rate than the actual database used.

2.1.3. Frameworked database loaders and filters

DBToolkit needs to be able to load a variety of sequence database formats in order to be widely applicable. However, as one can not foresee nor support all possible database formats out-of-the-box, it is reasonable to develop DBToolkit in such a way that interpreters for additional database formats can be added easily at a later time by third-party programmers. From the perspective of the user it is of course also important that DBToolkit can quickly and automatically recognize known database formats, and this for both built-in as well as added interpreters.

[22] The identification ratio is the number of identified spectra divided by the total number of presented spectra.

Additionally, one of the typical tasks in processing sequence databases is the filtering of a database to obtain a certain subset of entries. Examples of these are the human-derived proteins in SWISS-PROT or the mouse proteins in the NCBI non-redundant database. Of course, the available filtering options depend heavily on the actual (meta-)information present in the database and are thus format-specific. Finally, it should also be easy to define and integrate new filtering capabilities by third-party developers, again for both built-in database formats as well as for added ones.
In order to achieve the above set of goals, the core design of DBToolkit relies on two interconnected frameworks for database loading and filtering.
The filtering framework performs an ostensibly straightforward task. It should enable a picky data consumer to ask whether a given database entry passes or fails the filter. In DBToolkit the main interface of this framework is the *Filter* class. It defines two methods to which any filter implementation must adhere. Each method accepts a database entry in a certain format and returns a boolean to indicate whether the entry passes this filter. To enable easy addition of custom-built filters, the filter-loading mechanism is completely dynamic. This means that filter classes are only loaded when they are explicitly requested. As such, it becomes possible to define known filters descriptively (in a text file) instead of declaratively (inside compiled code). Adding a new filter to DBToolkit is therefore reduced to writing a suitable implementation for the *Filter* interface, and describing the resulting class in the correct text file. No recompilation of the DBToolkit sources is required[23].
The framework for database loading centers round the *DBLoader* interface. There are two abstract subclasses available for this interface, *DefaultDBLoader* and *ZippedDBLoader*. The former focuses on loading data straight from the database text files, while the second will load the same data from zipped or GZIPped text files (since sequence databases can be very large, this latter functionality is often useful to limit the strain on disk space). More interestingly however, both abstract super classes implement the following methods that were defined on the *DBLoader* interface: nextFilteredProtein(Filter aFilter) and nextFilteredRawEntry(Filter aFilter). It is of note that the effective implementation of these methods can be provided at this abstract level due to the successful abstraction of the filtering functionality. These two methods thus define the complete intersection between the two frameworks.
The frameworks and their connection are represented as Unified Modelling Language (UML) diagrams[24] in figure 33.

[23] It is not even required to have the DBToolkit sources at all.
[24] This UML diagram is not a conceptual model; it is automatically generated on the basis of the actual DBToolkit code itself.

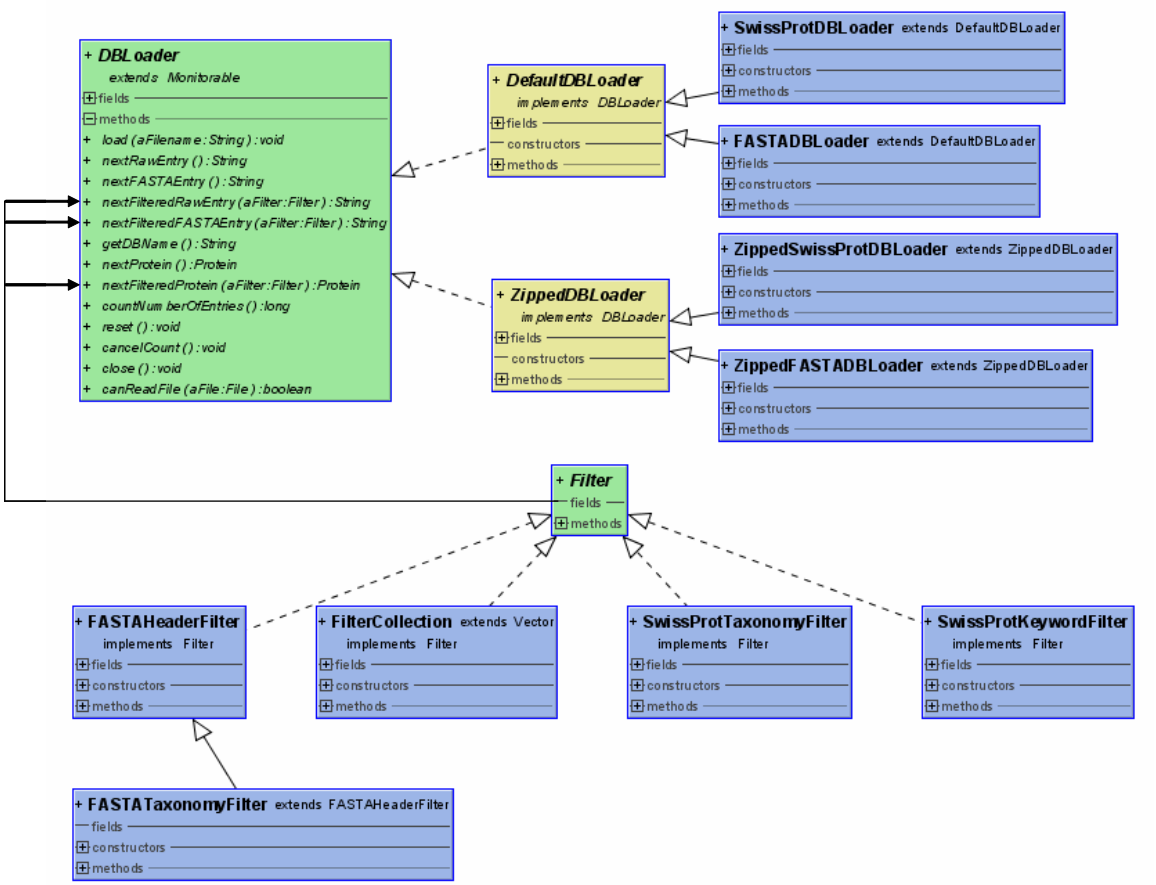


Figure 33: DBLoader and Filter frameworks and their connection.

The *DBLoader* framework also relies on a text file that lists the available database loaders and is thus fully descriptive in nature as well. Addition of a database loader therefore requires only an implementation of *DBLoader* (usually through extension of one of the abstract ancestors) and an additional line in the correct text file detailing the fully qualified class name for the loader. Again, no recompilation of DBToolkit is necessary at any point.

Finally, the *DBLoader* interface also defines the canReadFile(File aFile) method, which is the keystone for automatically detecting the correct *DBLoader* implementation for a given database file. The corresponding factory (*DBLoaderLoader*) can therefore sequentially present a database file to all registered *DBLoader* implementations, requesting whether the implementation recognizes the file formatting. It is of note that this mechanism represents an implementation of a simplified version of a command pattern and allows the automatic recognition of known database formats irrespective of whether a *DBLoader* implementation was packaged with DBToolkit or separately developed and added by a third party.

2.1.4. License and availability

The DBToolkit software has been released as open source software under the GNU[25] General Public License (GPL) (http://www.gnu.org/licenses/licenses.html#GPL) and is freely available in both source and binary versions from http://genesis.UGent.be/dbtoolkit.

2.1.5. Publication

[25] Interestingly, GNU is a recursive acronym for 'Gnu is Not Unix'.

BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 17 2005, pages 3584–3585
doi:10.1093/bioinformatics/bti588

Databases and ontologies

DBToolkit: processing protein databases for peptide-centric proteomics

Lennart Martens*, Joël Vandekerckhove and Kris Gevaert

Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University and Flanders Interuniversity Institute for Biotechnology (VIB09), A. Baertsoenkaai 3, B-9000 Ghent, Belgium

Received on June 10, 2005; accepted on July 14, 2005
Advance Access publication July 19, 2005

ABSTRACT

Summary: DBToolkit is a user-friendly, easily extensible tool that allows the processing of protein sequence databases to peptide-centric sequence databases. This processing is primarily aimed at enhancing the useful information content of these databases for use as optimized search spaces for efficient identification of peptide fragmentation spectra obtained by mass spectrometry. In addition, DBToolkit can be used to reliably solve a range of other typical tasks in processing sequence databases.

Availability: DBToolkit is open source under the GNU GPL license. The source code, full user and developer documentation and cross-platform binaries are freely downloadable from the project website at http://genesis.UGent.be/dbtoolkit/

Contact: lennart.martens@UGent.be

INTRODUCTION

As the tool of choice in present-day high-throughput proteomics, mass spectrometry has evolved substantially over the last years. The classical approach of two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) (O'Farrell, 1975) requires merely a mass measurement of the peptides generated from an enzymatic digest of an isolated protein (Cottrell, 1994). A refinement of this approach uses fragmentation spectra of a few peptides as additional information for the identification of the original protein. In most recent so-called gel-free techniques (e.g. as reviewed by Zhang *et al*., 2004 and Gevaert *et al*., 2005) however, mass spectrometers must be able to generate high-quality fragmentation spectra from extremely complex peptide mixtures, obtained following proteolytic digestion of an unfractionated proteome of a cell or tissue. These peptide-centric methods were primarily developed to deal with the inherent shortcomings of classical 2D-PAGE techniques and allow for a greater coverage of the proteome while simultaneously increasing the sensitivity of the analysis (Aebersold and Mann, 2003). The peptide-centric technologies have driven the exchange of the protein for the peptide as the basic unit in proteomics research (Aebersold and Mann, 2003; Kearney and Thibault, 2003).

Sequence databases like SWISS-PROT, IPI and the NCBI non-redundant database remain protein-based however. Since this discrepancy between the respective fundamental units can lead to a loss of highly interesting identifications, we developed the DBToolkit suite of software tools to allow the conversion of protein sequence databases into peptide sequence databases.

*To whom correspondence should be addressed.

APPLICATION FUNCTIONALITIES

The software can recognize FASTA and EMBL formatted databases out of the box, with UniProt and IPI the most prominent examples of the latter. It is also extremely easy for developers to include automatic recognition of different database formats as detailed below.

DBToolkit can perform various types of processing on sequence databases. Of course, simple *in silico* enzymatic digests using a variety of predefined enzymes or user-added enzymes are possible as well as database concatenation and FASTA output of differently formatted databases. The enzymatic digest even allows for 'dual specificity' enzymes that generate peptides for which the aminoterminus (N-terminus) is the result of a different cleavage pattern than the carboxyterminus (C-terminus). In addition, it is also possible to filter databases (the exact filtering options depend on the database format loaded) and to limit output to sequences in a certain mass range. Additional filters by other developers are also readily included in the software (see below). The three most powerful functions of DBToolkit however, are sequence-based filtering through a simple query language, N-terminal or C-terminal ragging (optionally truncating sequences in the process) and sequence-based redundancy clearing. The ragging process creates a series of subsequences for each 'mothersequence' where in each n-th subsequence, the first $n-1$ residues have been removed from the N-terminal or C-terminal side, respectively.

These functions are readily applied serially to achieve compound results such as a non-redundant, N-terminally ragged subset of a trypsin digest of the *Homo sapiens* entries in the UniProt database, all of which have a mass between 600 and 4000 Da.

Several applications for these processed databases are outlined below.

APPLICATION DESIGN

DBToolkit is completely written in the Java programming language and its only requirement is a Java runtime environment 1.3 or above.

The suite consists of both an intuitive graphical user interface presenting the user with interactive controls to all processing steps, and an equivalent set of command-line tools for straightforward automation of the processing steps through simple scripting. This latter functionality has allowed us to tie different processing steps in with the automatic database updating of Mascot (http://www.matrixscience.com) for the most popular sequence databases, creating multiple derived databases overnight.

DBToolkit was designed from the start to be easily extensible. The use of robust frameworking allows the addition of novel database loaders or filters without requiring recompilation.

Full user and developer documentation for the suite is available from the project website, along with the cross-platform binaries and CVS repository coordinates.

DISCUSSION

We have applied DBToolkit in the lab for numerous purposes, most notably the generation of specialized databases for use as searchbases for protein identification in Mascot. One approach used ragged, non-redundant peptide databases to increase the number of identified spectra in an N-terminal COFRADIC experiment with ∼40% (Gevaert *et al.*, 2003). Interestingly, most of the peptides identified only in the ragged databases corresponded to the novel N-termini of their progenitor proteins after *in vivo* processing (e.g. the N-termini of nuclear-encoded proteins that are imported into mitochondria and lost their transit peptide). Since these processing sites typically did not conform to standard tryptic sites, they were absent from searches solely performed in the original sequence databases. Another application has been found in picking up peptides from apoptose substrates, yielding the exact cleavage location in those proteins. For this we created non-redundant, enzymatically digested peptide databases using a bifunctional enzyme that created peptides with an N-terminus derived from caspase activity (i.e. consensus cleavage C-terminal to aspartic acid) and a C-terminus derived from trypsin activity. In this way, a large number of caspase cleavage sites have been confirmed and many tentative new sites have been found that would otherwise have eluded identification (unpublished data). A third application centers on the a priori calculation of the potential success a certain COFRADIC procedure could have by rapidly creating non-redundant, comprehensive lists of all detectable peptides containing a specified amino acid. Note that this functionality can be applied to any peptide-centric proteomics approach that can select for sequences by their aminoacid content (see Zhang *et al*., 2004 and Gevaert *et al*., 2005 for an overview of these techniques).

DBToolkit has proven to be a highly versatile yet very simple tool for routine tasks in sequence database processing. Furthermore, as the applicability and popularity of peptide-centric proteomics experiments expands further, DBToolkit can perform the essential task of complementing proven, probabilistic protein identification software like Mascot with peptide-centric search databases, optimized for the specific conditions and requirements of the research.

ACKNOWLEDGEMENTS

L.M. would like to thank An Staes, Evy Timmerman, Petra Van Damme, Grégoire Thomas and Luc Krols for their useful suggestions and comments on the DBToolkit software during its development phase. K.G. is a Postdoctoral Fellow and L.M. a Research Assistant of the Fund for Scientific Research, Flanders (Belgium) (FWO, Vlaanderen). The project was supported by research grants from the Fund for Scientific Research, Flanders (Belgium) (project number G.0008.03), the Inter University Attraction Poles (IUAP, project number P5/05), the GBOU-research initiative (project number 20204) of the Flanders Institute of Science and Technology (IWT) and the European Union Interaction Proteome (6th Framework Program). Funding to pay the Open Access publication charges for this article was provided by Ghent University, Belgium.

Conflict of Interest: none declared.

REFERENCES

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

Cottrell,J.S. (1994) Protein identification by peptide mass fingerprinting. *Pept. Res.*, **7**, 115–124.

Gevaert,K. *et al.* (2003) Exploring proteomes and analyzing protein processing bymass spectrometric identification of sorted N-terminal peptides. *Nat. Biotechnol.*, **21**, 566–569.

Gevaert,K. *et al.* (2005) Diagonal reverse-phase chromatography applications in peptide-centric proteomics; ahead of catalogue-omics? *Anal. Biochem.*, in press.

Kearney,P. and Thibault,P. (2003) Bioinformatics meets proteomics—bridging the gap between mass spectrometry data analysis and cell biology. *J. Bioinform. Comput. Biol.*, **1**, 183–200.

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

Zhang,H. *et al.* (2004) Chemical probes and tandem mass spectrometry: a strategy for the quantitative analysis of proteomes and subproteomes. *Curr. Opin. Chem. Biol.*, **8**, 66–75.

2.1.6. Published applications

2.1.6.1. N-terminal proteome of unstimulated human blood platelets

Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides

Kris Gevaert, Marc Goethals, Lennart Martens, Jozef Van Damme, An Staes, Grégoire R. Thomas and Joël Vandekerckhove

Published online 31 March 2003; doi:10.1038/nbt810

Current non-gel techniques for analyzing proteomes rely heavily on mass spectrometric analysis of enzymatically digested protein mixtures. Prior to analysis, a highly complex peptide mixture is either separated on a multidimensional chromatographic system[1,2] or it is first reduced in complexity by isolating sets of representative peptides[3–8]. Recently, we developed a peptide isolation procedure based on diagonal electrophoresis[9] and diagonal chromatography[10]. We call it combined fractional diagonal chromatography (COFRADIC). In previous experiments, we used COFRADIC to identify more than 800 *Escherichia coli* proteins by tandem mass spectrometric (MS/MS) analysis of isolated methionine-containing peptides[11]. Here, we describe a diagonal method to isolate N-terminal peptides. This reduces the complexity of the peptide sample, because each protein has one N terminus and is thus represented by only one peptide. In this new procedure, free amino groups in proteins are first blocked by acetylation[12] and then digested with trypsin. After reverse-phase (RP) chromatographic fractionation of the generated peptide mixture, internal peptides are blocked using 2,4,6-trinitrobenzenesulfonic acid (TNBS)[13,14]; they display a strong hydrophobic shift and therefore segregate from the unaltered N-terminal peptides during a second identical separation step. N-terminal peptides can thereby be specifically collected for further liquid chromatography (LC)-MS/MS analysis. Omitting the acetylation step results in the isolation of non-lysine-containing N-terminal peptides from *in vivo* blocked proteins.

We used this technique to identify 264 proteins and 78 *in vivo*–acetylated proteins in a cytosolic and membrane skeleton fraction of human thrombocytes. In addition to showing that this method can be used for gel-free proteomics, we demonstrate that it allows one to examine N-terminal protein processing such as removal of signal sequences and modifications. A general scheme depicting the procedure for sorting N-terminal peptides is shown in Figure 1. At the protein level, cysteines are alkylated with iodoacetamide, and free primary amines are blocked by acetylation. Upon trypsin digestion, two types of peptides are generated: internal peptides with a free α-amino group and blocked N-terminal peptides. These peptides are fractionated by RP–high-performance (HP) LC, typically in 12 fractions (primary run, Fig. 2A). The dried fractionated peptides are redissolved in an appropriate buffer for reaction with TNBS. Only internal peptides react with TNBS to form very hydrophobic trinitrophenyl-peptides (TNP-peptides), whereas blocked N-terminal peptides are not affected. On average, this TNBS modification reaction proceeds to more than 98% completion. Each TNBS-treated primary fraction is separately rerun on the same column and under conditions identical to those for the primary run. The TNP-labeled internal peptides now shift to later elution times and separate from the unaltered N-terminal peptides, which do not shift and can be easily collected for further analysis (Fig. 2B). The secondary run and analysis step is repeated for each TNBS-modified fraction.

With this procedure we analyzed the proteome of a cytosolic and membrane skeleton fraction of human thrombocytes[15]. Sorted N-terminal peptides in all fractions were analyzed using 96 LC-MS/MS runs, during which 5,640 collision-induced dissociation (CID)

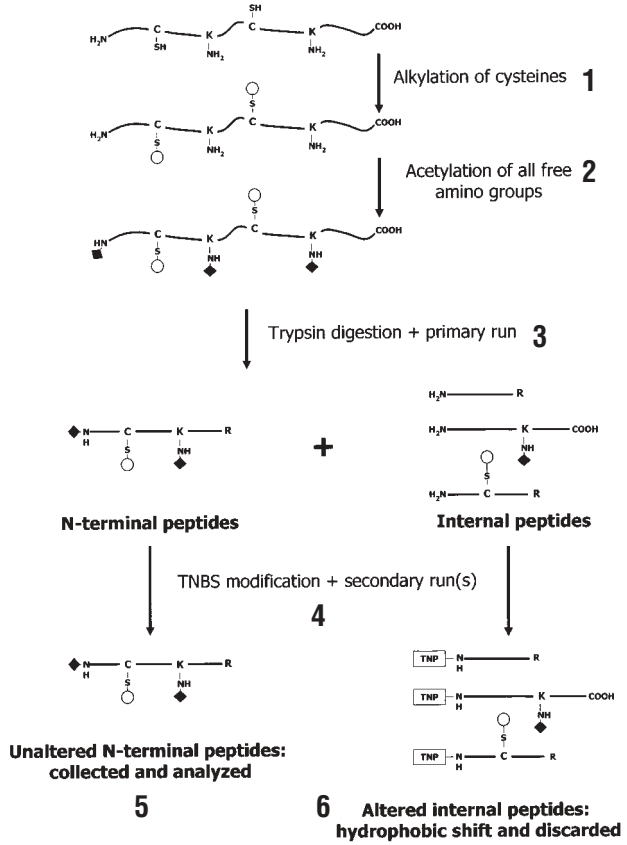


Figure 1. Scheme summarizing the chemistry and chromatographic steps during N-terminal peptide sorting. (1) All protein-cysteine residues are first alkylated using iodoacetamide (open circles). (2) Then, all free amines (α- and ε-amines) are acetylated (indicated by filled diamonds) and the proteins are digested with trypsin (3), which will now only cleave C-terminal to arginine residues. As shown, this creates two types of peptides: N-terminal peptides with a blocked N terminus and internal peptides with a free N terminus. Following the primary RP-HPLC fractionation of the generated peptide mixture, all peptides present in one HPLC fraction are treated with TNBS (4). Only the internal peptides are altered to trinitrophenyl-peptides (indicated by TNP in open boxes), which have become more hydrophobic and will thus shift out of their original position during the secondary chromatographic run (5), which is identical to the first RP-HPLC separation. These internal peptides are discarded for further analysis. However, N-terminal peptides are unaltered by the TNBS reaction, thus elute at exactly the same time interval as during the first run and can be specifically collected and analyzed further by LC-MS/MS (6).

Department of Medical Protein Research, Flanders Interuniversity Institute for Biotechnology, Department of Biochemistry, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium.
**Corresponding author (kris.gevaert@rug.ac.be).*

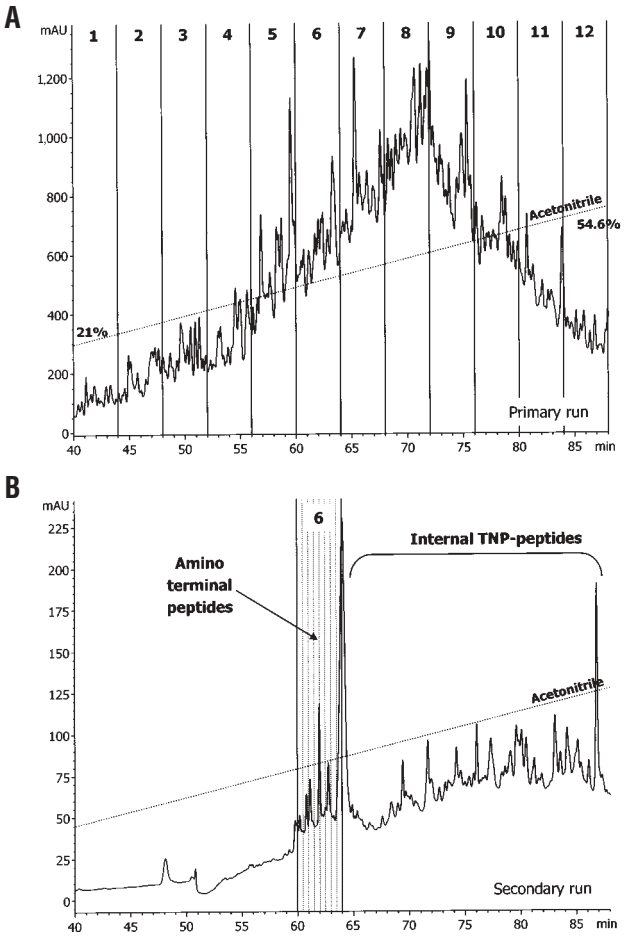


Figure 2. Peptide elution profiles during the primary and secondary run of N-terminal peptide sorting. (A) Chromatographic fractionation of a tryptic digest of a human thrombocyte protein fraction. During this primary run, a total of 12 fractions are collected and treated with TNBS to block the free amines. When the TNBS-treated fractions are separately rerun on the same column and under identical conditions, the internal (trinitrophenyl) peptides shift to later elution times (hydrophobic shift), whereas the unaltered N-terminal peptides elute within the same time interval and can be collected in a number of secondary fractions (only shown for primary fraction 6 in B).

spectra were generated. Because proteins are variably processed at their N termini, their exact *in vivo* N-terminal boundaries are often not known. This poses a problem for peptide identification when conventional protease-based MS/MS-based peptide identification routes are followed. Therefore, we created databases containing sequentially N-terminally trimmed Arg-C-type peptides. Using these peptide-centric databases in combination with established database searching tools, the number of peptides identified increased by 50%. In total, 1,247 CID spectra were unambiguously assigned to peptides using MASCOT[16], implying that 22.83% of all MS/MS spectra obtained could be positively identified. This is somewhat less than the proportion (34.48%) identified in the previously published *E. coli* study[11]; however, in that study, all the open reading frames were publicly available and could thus be searched using MS/MS data. Of the spectra we identified, 920 (73.77%) correspond to 305 different acetylated peptides, from which 264 unique human proteins were identified (see Supplementary Table 1 online). Thus, on average, each protein was identified by one peptide, clearly indicating the reduction in the complexity of the analyte mixture. The proteins were ranked according to the frequency at which their N-terminal peptides were identified (see Supplementary Table 2 online). This ranking can be considered as a semiquantitative measure of their abundance in the sample (the 50 most abundant proteins are shown).

Proline and pyroglutamate residues seem to react slowly or not at all with TNBS. Peptides with such N-terminal residues do not show a shift in our system and are isolated together with the blocked N-terminal peptides (see Supplementary Table 1 online). Some proline-peptides clearly constitute the extreme N terminus of proteins, whereas others may have been derived from internal cleavage, either *in situ* or during protein preparation. An example is transgelin (P37802), for which both the intact acetyl-ANRGPAYGLSR and truncated PAYGLSR peptides were identified (see Supplementary Table 1 online). Pyroglutamate-peptides represent about 10% of the selected peptides and therefore do not hamper further LC-MS/MS analysis of sorted N-terminal peptides, but rather may provide additional information when N-terminal peptide analysis fails. This is illustrated by nonmuscle myosin heavy chain (P35579), where such peptides are frequently encountered, whereas the N terminus is not found (see Supplementary Table 1 online). Next to these 'expected contaminants,' we identified 74 internal tryptic peptides, derived from 47 abundant proteins. The presence of these unsorted internal peptides is due to incomplete reaction of some abundant peptides with TNBS. This number can be reduced with an additional coupling step, which was omitted here in order to limit losses. The number of internal peptides is very low compared to the number of N termini discovered. For instance, an *in silico* Arg-C digestion of the human proteome reveals the presence of 17.5 internal peptides per N-terminal peptide. With 305 different N termini identified, we would encounter 5,337 (17.5×305) internal peptides if no sorting were done. The 74 internal peptides detected show that our N-terminal selection procedure had an overall efficiency of more than 98%.

Excluding the amino acetylation step (Fig. 1) automatically leads to the isolation of the N-terminal peptides of *in vivo* N-blocked proteins. These peptides should not contain lysine, as the latter will also react in its unprotected form with TNBS. Thus, here, we will only sort for arginine-containing peptides with a blocked N terminus, representing approximately half of the *in vivo*–blocked proteins. In our platelet fraction, this amounts to 78 acetylated proteins (see Supplementary Table 3 online).

Using our procedure, established or predicted N-terminal protein processing can easily be verified. For example, the signal peptides of two identified putative mitochondrial proteins, malic enzyme and dihydroorotate dehydrogenase, have been predicted by homology. Here we found that the malic enzyme 2 (NAD-dependent malic enzyme) starts at residue 19 rather than at amino acid 21 (http://us.expasy.org/cgi-bin/niceprot.pl?P23368) and dihydroorotate dehydrogenase starts at residue 28 rather than at amino acid 11 (http://us.expasy.org/cgi-bin/niceprot.pl?Q02127) (see Supplementary Table 4 online). Identified putative membrane proteins (see Supplementary Table 5 online) segregate in three groups: those with an N-terminal signal sequence (type I), those lacking a pre-sequence, and those that are cleaved internally. *In vivo*–blocked proteins fall into four categories (see Supplementary Table 6 online): those carrying an acetylated initiator methionine, generally followed by an acidic residue; those with the initiator methionine removed and with the second residue acetylated; those from which a small number of N-terminal residues are removed before acetylation; and those that are cleaved internally and re-acetylated. An interesting representative of the latter group is the hitherto undiscovered form of a truncated form of actin. Normally, actins are acetylated on position three; this new variant starts at residue 29 (see Supplementary Table 6 online).

Various N-terminal modifications can also be considered while searching the protein databases. When N-terminal formylation was considered, a variant of the macrophage migration inhibitory factor (MIF, NCBI GenBank accession no. 5542179) was found to be formylated (MIF_{PAM}), whereas the protein lacking this alanine residue

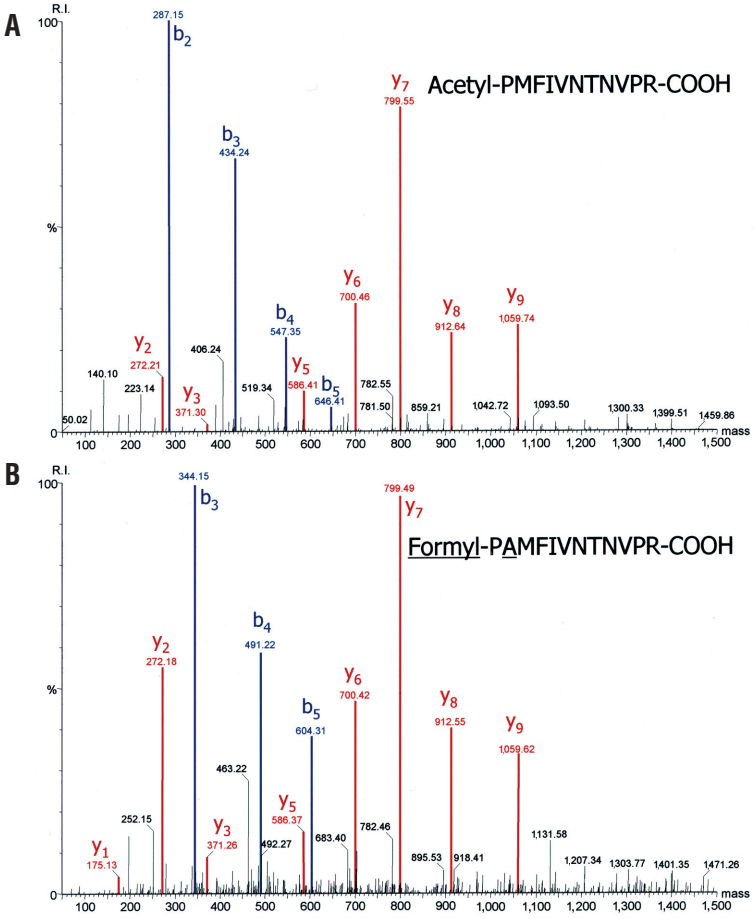


Figure 3. Acetylation and formylation of the macrophage migration inhibitory factor. Q-TOF MS/MS spectra of the acetylated N-terminal peptide (doubly charged ion of *m/z* = 673.43) belonging to the macrophage migration inhibitory factor (P14174) (A) and a formylated N-terminal peptide (*m/z* = 701.80) of the isoform carrying an alanine inserted between Pro-1 and Met-2 (NCBI GenBank accession no. 5542179) (B). Although the pattern of the observed y-ions (shown in red) is very alike, the b-ion series (shown in blue), carrying the differently modified N terminus, differs significantly. The differences correspond to the N-terminal acetyl/formyl exchange.

(MIF$_{PM}$) yielded an acetylated peptide (Fig. 3). Interestingly, the corresponding MIF$_{PM}$ peptide is present in the list of *in vivo*–blocked proteins (see Supplementary Table 3 online). Therefore, we conclude that the MIF$_{PM}$ variant was only acetylated during the first step of our procedure (Fig. 1). Thus, *in vivo*, MIF$_{PM}$ has a free N-terminal proline, whereas MIF$_{PAM}$ is formylated.

Because N-terminal peptide mixtures are much less complex than an unfractionated mixture[1,2] or mixtures of selected Met-containing peptides[11], our approach may lead the way to a proteomics approach in which proteins are identified by the mass of their N-terminal peptides only. Such an approach should be feasible using highly accurate mass spectrometers measuring with errors less than 1 ppm[17]. This will significantly accelerate proteomics and bring the field one step closer to high-throughput analysis for diagnostics and drug discovery.

Experimental protocol

Sorting of N-terminal peptides. A cytosolic and membrane skeleton fraction prepared from 10^9 human thrombocytes[15] (obtained from the Red Cross Blood Transfusion Centre Oost-Vlaanderen, Ghent, Belgium) was dissolved in 4 M guanidium chloride (Fluka Chemie GmbH, Buchs, Switzerland) in 0.1 M sodium phosphate buffer (pH 7.0). Proteins were reduced for 90 min in 0.02% tributylphosphine (Fluka) and alkylated for 30 min at 37 °C in 5 mM iodoacetamide (Fluka). About 10 nmol of protein material was desalted on a NAP-5 column (Amersham Pharmacia Biotech, Uppsala, Sweden) in 1 ml of 1 M guanidinium chloride in 50 mM sodium phosphate buffer (pH 8.0.) and reduced to half its volume by vacuum drying.

Free amines were acetylated in 10 mM sulfo-*N*-hydroxysuccinimide acetate (Perbio, Helsingborg, Sweden) for 90 min at 30 °C. This step was omitted for the isolation of lysine-free N termini of *in vivo*–acetylated proteins. Partial acetylation of serine and threonine was reversed by adding 1 µl of hydroxylamine (Fluka) to the protein mixture. This mixture was desalted on a NAP-5 column in 1 ml of 0.25 M guanidinium chloride in 50 mM Tris-HCl (pH 7.9) and reduced to 0.5 ml by vacuum drying. It was then boiled for 5 min and digested overnight at 37 °C with 10 µg of sequencing-grade modified trypsin (Promega Corporation, Madison, WI).

The peptide mixture was separated on a RP-HPLC column (2.1 mm internal diameter × 150 mm 300SB-C18 column; Zorbax, Agilent Technologies, Waldbronn, Germany) on an Agilent 1100 Series HPLC system. Following a 10 min wash with solvent A (98:2 (vol/vol) water/acetonitrile in 0.1% trifluoroacetic acid (TFA), both Baker HPLC analyzed, Mallinckrodt Baker B.V., Deventer, The Netherlands), a linear gradient to 100% solvent B (30 parts 0.1% (vol/vol) TFA in water/70 parts acetonitrile) was applied over 100 min with a flow rate of 80 µl/min. Peptides eluting between 40 and 88 min were collected in 12 fractions of 4 min each in a 96 microwell plate.

Three consecutive programs were written for the Agilent well plate sampler for the TNBS modification of peptides. First, the dried peptides were redissolved in 50 µl of 50 mM sodium borate (pH 9.5). Then, 2 µmol of dried TNBS (stored in two separate vials) were redissolved in 200 µl of 50 mM sodium borate (pH 9.5), of which 15 µl was transferred to each sample well. The peptides were incubated with TNBS for 55 min at 37 °C. This reaction was repeated once, after which the samples were dried. The modification procedure, including the two reaction steps, was repeated once to assure nearly quantitative TNBS modification.

The samples were dried a final time before the secondary HPLC runs. Each secondary run began with an injector program in which the sample was redissolved in 70 µl 0.5% TFA, of which 65 µl was loaded onto the column. The same solvent gradient was used as in the primary run. N-terminal peptides were collected in the same 4 min time interval as used during the primary run in eight subfractions of 30 s (or 40 µl). We thus collected 96 subfractions for subsequent LC-MS/MS analysis[11]. Briefly, after drying, the peptides were redissolved in 20 µl of 0.1% formic acid in water and 10 µl of this solution was separated by nano-RP-HPLC connected to a Q-TOF1 mass spectrometer (Micromass UK Ltd., Cheshire, UK). Mass spectra obtained by automated LC-MS/MS analysis were used to identify the corresponding peptides using MASCOT[16] (see Supplementary Experimental Protocol and Supplementary Figure 1 online).

Note: Supplementary information is available on the Nature Biotechnology website.

Acknowledgments

K.G. is a postdoctoral fellow and L.M. a research assistant of the Fund for Scientific Research–Flanders (Belgium) (F.W.O.–Vlaanderen). The project was further supported by the Concerted Research Actions (GOA) of the Flemish Community, the Inter University Attraction Poles (IUAP), and the GBOU-research initiative of the Flanders Institute of Science and Technology (IWT).

Competing interests statement

The authors declare that they have no competing financial interests.

Received 23 October 2002; accepted 30 January 2003

1. Shen, Y. *et al.* High-throughput proteomics using high-efficiency multiple-capillary liquid chromatography with on-line high-performance ESI FTICR mass spectrometry. *Anal. Chem.* **73**, 3011–3021 (2001).
2. Washburn, M.P., Wolters, D. & Yates III, J.R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–248 (2001).
3. Gygi, S.P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).
4. Geng, M., Ji, J. & Regnier, F.E. Signature-peptide approach to detecting proteins in complex mixtures. *J. Chromatogr. A.* **870**, 295–313 (2000).
5. Oda, Y., Nagasu, T. & Chait, B.T. Enrichment analysis of phosphorylated proteins

as a tool for probing the phosphoproteome. *Nat. Biotechnol.* **19**, 379–382 (2001).
6. Wang, S. & Regnier, F.E. Proteomics based on selecting and quantifying cysteine-containing peptides by covalent chromatography. *J. Chromatogr. A.* **924**, 345–357 (2001).
7. Zhou, H., Watts, J.D. & Aebersold, R. A systematic approach to the analysis of protein phosphorylation. *Nat. Biotechnol.* **19**, 375–378 (2001).
8. Zhou, H., Ranish, J.A., Watts, J.D. & Aebersold, R. Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nat. Biotechnol.* **20**, 512–515 (2002).
9. Brown, J.R. & Hartley, B.S. Location of disulphide bridges by diagonal paper electrophoresis. The disulphide bridges of bovine chymotrypsinogen A. *Biochem. J.* **101**, 214–228 (1966).
10. Cruickshank, W.H., Malchy, B.L. & Kaplan, H. Diagonal chromatography for the selective purification of tyrosyl peptides. *Can. J. Biochem.* **52**, 1013–1017 (1974).
11. Gevaert, K. *et al.* Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis: identification of more than 800 *Escherichia coli* proteins. *Mol. Cell. Proteomics* **1**, 896–903 (2002).
12. Yumoto, N. & Tokushige, M. Acetylation-induced alteration of catalytic and regulatory properties of aspartase. *Biochim. Biophys. Acta* **749**, 101–115 (1983).
13. Okuyama, T. & Satake, K. On the preparation and properties of 2,4,6-trinitrophenyl-amino acids and –peptides. *J. Biochem. (Tokyo)* **47**, 454 (1960).
14. Fields, R. The rapid determination of amino groups with TNBS. *Methods Enzymol.* **25**, 464–468 (1972).
15. Fox, J.E. Identification of actin-binding protein as the protein linking the membrane skeleton to glycoproteins on platelet plasma membranes. *J. Biol. Chem.* **260**, 11970–11977 (1985).
16. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
17. O'Connor, P.B. & Costello, C.E. Internal calibration on adjacent samples (InCAS) with Fourier transform mass spectrometry. *Anal. Chem.* **72**, 5881–5885 (2000).

Continuous high-titer HIV-1 vector production

Yasuhiro Ikeda[1], Yasuhiro Takeuchi[1], Francisco Martin[1], Francois-Loic Cosset[2], Kyriacos Mitrophanous[3], and Mary Collins[1]*

Published online 7 April 2003; doi:10.1038/nbt815

Human immunodeficiency virus type 1 (HIV-1)–based vectors are currently made by transient transfection, or using packaging cell lines in which expression of HIV-1 Gag and Pol proteins is induced[1–3]. Continuous vector production by cells in which HIV-1 Gag-Pol is stably expressed would allow rapid and reproducible generation of large vector batches. However, attempts to make stable HIV-1 packaging cells by transfection of plasmids encoding HIV-1 Gag-Pol have resulted in cells which secrete only low levels of p24 antigen (20–80 ng/ml)[4–6], possibly because of the cytotoxicity of HIV-1 protease[7]. Infection of cells with HIV-1 can result in stable virus production[8]; cell clones that produce up to 1,000 ng/ml secreted p24 antigen have been described[9]. Here we report that expression of HIV-1 Gag-Pol by a murine leukemia virus (MLV) vector allows constitutive, long-term, high-level (up to 850 ng/ml p24) expression of HIV-1 Gag. Stable packaging cells were constructed using codon-optimized HIV-1 Gag-Pol[10] and envelope proteins of gammaretroviruses; these producer cells could make up to 10^7 293T infectious units (i.u.)/ml (20 293T i.u./cell/day) for at least three months in culture.

The vectors pCNC-GPRT and pCNC-SYNGP (Fig. 1A) were either transfected into HeLa or HT1080 cells, or first packaged into MLV virions using a transient MLV packaging system[11] and then used to infect HeLa, HT1080, or 293T cells. Transfected or infected HeLa and HT1080 cells were selected in G418; infected 293T cells were cloned by limiting dilution, as the cells are already G418 resistant. Each clone was then analyzed for HIV-1 p24 capsid (CA) expression by immunofluorescence microscopy. Compared to transfection, infection with either MLV vector generated a higher frequency of clones expressing HIV-1 p24 (Table 1A). In all cell lines, infection with the codon-optimized CNC-SYNGP vector generated a higher frequency of positive clones than the CNC-GPRT vector (Table 1A).

To determine which of the cell lines would make the best stable packaging cells, we measured p24 secretion. The infected HeLa cells secreted more p24 than the transfected cells (Fig. 1B). The amount of p24 secreted by 293T cells infected with wild-type HIV-1 Gag-Pol could be increased by using the MLV vector CNC-Rev to express additional Rev, which controls nuclear export of HIV-1 RNA (Fig. 1B). This suggests that pCNC-GPRT expresses a suboptimal level of Rev. Clones of 293T and HT1080 cells infected with CNC-SYNGP produced the most p24 (Fig. 1B). Greater p24 secretion by infected cells, compared to transfected cells, resulted from higher RNA expression. Infected HeLa cells (clone B, Fig. 1B) expressed considerably more HIV-1 Gag-Pol RNA than the transfected cells (clone A) (Fig. 1C). The predominant RNA species was the correct size for the transcript driven by the cytomegalovirus immediate early (CMV) promoter. This higher RNA expression in the infected cells did not result from a higher number of CNC-GPRT copies (Fig. 1C). 293T cells secreted more p24 than HeLa cells for a given level of RNA expression (Fig. 1B,C). Immunoblot analysis of cell lysates and cell supernatants showed that HeLa cells retain higher levels of HIV-1 Gag precursors within the cell (data not shown). Because of this higher p24 secretion and the efficiency of transfection, we pursued packaging-cell construction with HT1080 and 293T cells.

To generate stable packaging cells, we chose to express envelopes of gammaretroviruses because they are not cytotoxic, can produce relatively high-titer pseudotypes of HIV-1, and have been used in clinical and preclinical gene therapy[12–14]. The 293T clone expressing GPRT and Rev (GPRT1+R1 cells) and clones of the HT1080SYNGP1 or 293TSYNGP1 cells engineered to express the HIV-1 transactivator of transcription Tat, and Rev (HT-STAR and STAR) were transfected with envelopes of MLV 4070A (Ampho[15]), RD114 with an HIV protease site introduced at the R-peptide cleavage site (RDpro), or gibbon ape leukemia virus (GALV) with an MLV cytoplasmic tail (GALV+[16]). In each case a clone expressing a high level of envelope was chosen (1 clone from 12). Packaging of three HIV-1 genomes—the Rev-encoding pH7G[10], pHRSIN-CSGW[17], or its non-self-inactivating derivative pHV (Fig. 2A)—was compared. The STAR-Ampho cells produced the highest titer of virus, over 10^7 i.u./ml from a bulk population after infection by the HIV-1 genome (Table 1B). Both H7G and HV produced high titers, suggesting that the amount of Rev in the packaging cell was sufficient for vector production (Table 1B). Transfection and selection of STAR-Ampho cells also gave a high titer of virus, with clones producing 10^7 i.u./ml of self-inactivating vectors (Table 1B). Like transient lentiviral vector preparations, virus produced from the STAR cells could transduce target cells at high efficiency (Fig. 2B) and could infect 293T cells arrested by aphidicolin (data not shown). Transiently produced virus with vesicular stomatitis virus glycoprotein (VSV-G) envelope was more infectious per mg p24 than stably produced virus with gammaretroviral envelopes (Fig. 2B). This was a property of the VSV-G envelope, rather than the stable packaging cells, as we observed the same increased infectivity when transiently produced virus with VSV-G envelope was compared to transiently produced virus with

[1]*Department of Immunology and Molecular Pathology, Windeyer Institute, University College London, 46 Cleveland St., London W1T 4JF, UK.* [2]*Vectorologie Retrovirale et Therapie Genique, Ecole Normale Superieure de Lyon, Lyon, France.* [3]*Oxford BioMedica Limited, The Oxford Science Park, Oxford, UK.* **Corresponding author (mary.collins@ucl.ac.uk).*

2.1.6.2. Proteolytic processing by caspases in apoptotic Jurkat T-cells

Caspase-specific and nonspecific *in vivo* protein processing during Fas-induced apoptosis

Petra Van Damme, Lennart Martens, Jozef Van Damme, Koen Hugelier, An Staes, Joël Vandekerckhove & Kris Gevaert

We generated a comprehensive picture of protease substrates in anti-Fas–treated apoptotic human Jurkat T lymphocytes. We used combined fractional diagonal chromatography (COFRADIC) sorting of protein amino-terminal peptides coupled to oxygen-16 or oxygen-18 differential labeling. We identified protease substrates and located the exact cleavage sites within processed proteins. Our analysis yielded 1,834 protein identifications and located 93 cleavage sites in 71 proteins. Indirect evidence of apoptosis-specific cleavage within 21 additional proteins increased the total number of processed proteins to 92. Most cleavages were at caspase consensus sites; however, other cleavage specificities suggest activation of other proteases. We validated several new processing events by immunodetection and by an *in vitro* assay using recombinant caspases and synthetic peptides containing presumed cleavage sites. The spliceosome complex appeared a preferred target, as 14 of its members were processed. Differential isotopic labeling further revealed specific release of nucleosomal components from apoptotic nuclei.

The interest in proteases and their substrates is steadily growing because of the major roles they have in crucial molecular events such as protein maturation, enzyme regulation, protein localization, protein complex formation and stabilization. Their importance is illustrated by the fact that up to 1,200 human genes are estimated to encode proteases[1] and that the present version of the MEROPS protease database (http://merops.sanger.ac.uk/) contains 505 known or putative peptidases. In agreement with their biological relevance, altered protease expression or activation, and/or substrate proteolysis, may trigger human diseases like rheumatoid arthritis, cancer, neurodegenerative and cardiovascular diseases[2,3]. Consequently, proteases are recognized as important drug targets[4], orienting research and technologies toward a better understanding of degradation processes within protein networks and their resulting physiological effects.

A cellular process that is largely steered by proteases is apoptosis, a genetically programmed, morphologically and physiologically distinct form of cell death. The cysteine-dependent aspartate-specific proteases or caspases are the main players in the initiation and execution of apoptosis[5]. Caspase cleavage occurs at restricted sites and the resulting cleavage products are often stable in cells and potentially acquire alternative functions.

To obtain a better understanding of the cellular events associated with the generation of these apoptosis-specific products, it is important to have access to a comprehensive picture of the apoptotic degradome. A compilation of different studies carried out in various cells and organisms has so far resulted in the identification of close to 300 different caspase substrates[6]. Most of the compiled approaches identified potential apoptotic substrates, but seldom the exact cleavage sites[7–14].

With the idea to develop a more direct and general procedure to measure protein degradation, we developed a differential gel-free proteomics technique that specifically identifies the processed site(s) within the protease substrate. As such, it is the most direct way to study protein processing. We used the method to identify protease substrates and to locate their cleavage sites in Fas (CD95/Apo1)-stimulated[15] human Jurkat T lymphocytes. We identified 2,462 peptides, 93 of which were due to specific proteolysis of substrate proteins in apoptotic cells. We confirmed several results by western blotting and by *in vitro* peptide cleavage. These studies provide for the first time a comprehensive view of the apoptotic degradome in a model system, and the method described here can be applied to similar analyses in other biomedically relevant systems.

RESULTS

Peptide sorting and quantitative analysis

We induced apoptosis in human Jurkat cells by Fas antibody treatment for 24 h. At this time point, fluorescence activated cell sorting (FACS) analysis showed that 70% of the cells were in apoptotic phases while the activities of caspases 3 and 7 reached their highest values (data not shown). We prepared lysates from apoptotic and living cells for differential N-terminal peptide analysis. The latter is a modification of a previously published peptide sorting procedure, COFRADIC[16]. We mixed peptides from both digests (one ^{16}O-labeled and one ^{18}O-labeled)[17] in a one-to-one ratio (total peptide amount) and sorted them by N-terminal COFRADIC.

Different cleavage scenarios lead to a repertoire of sorted peptides (**Fig. 1**). As expected, the vast majority of these (over

Department of Medical Protein Research, Flanders Interuniversity Institute for Biotechnology, Department of Biochemistry, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium. Correspondence should be addressed to K.G. (kris.gevaert@ugent.be).

RECEIVED 10 JUNE; ACCEPTED 11 AUGUST; PUBLISHED ONLINE 22 SEPTEMBER 2005; DOI:10.1038/NMETH792

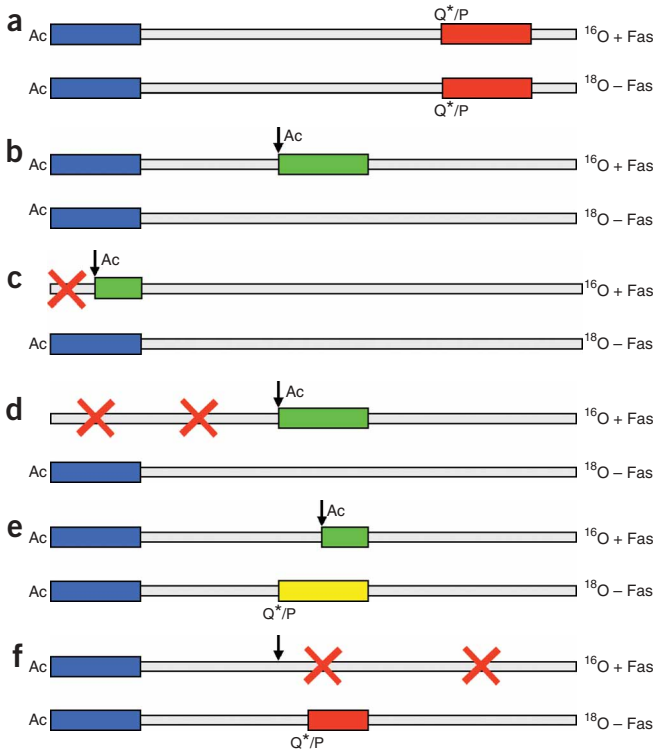
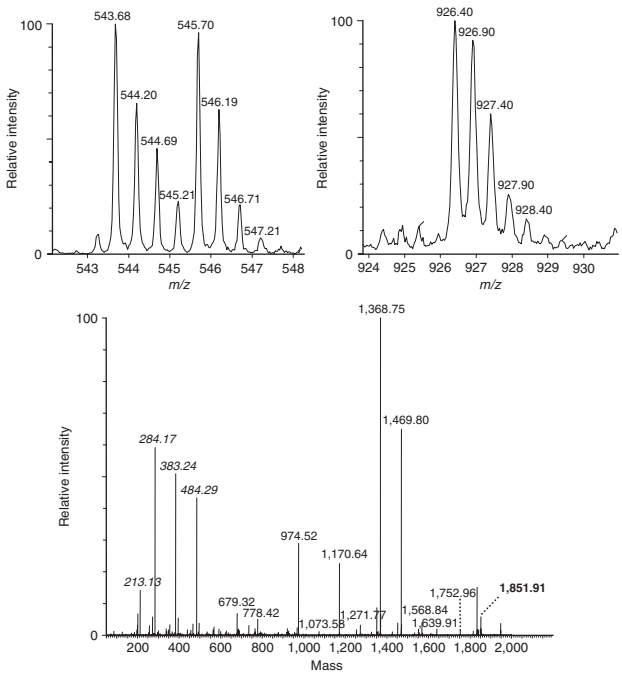


Figure 1 | Scenarios leading to different categories of N-blocked peptides in apoptotic lysates. (**a**) No cleavage in both samples. (**b**) The protein is cleaved in apoptotic cells: a ^{16}O-labeled α-N-acetylated peptide is generated. (**c**) Cleavage is very close to the N terminus: the short N-terminal peptide is lost. (**d**) The N-terminal part of the protein is degraded after cleavage. (**e**) The protein is cleaved at a site, fortuitously located within a Q* or P peptide: the ^{18}O variant is recovered; sometimes the new α-N-acetylated peptide is observed. (**f**) The C-terminal part of the protein is degraded after cleavage. Proteins are shown as long bars and cleavage sites are indicated with vertical arrows. Tryptic peptides from apoptotic lysates (+ Fas) are ^{16}O-labeled, and peptides from living control cells (− Fas) are ^{18}O-labelled. Colored bars represent peptides that were sorted by the N-terminal COFRADIC procedure (either α-N-acetylated (Ac) or start with PCA (Q*) or proline (P)): peptides at the extreme N termini of proteins (blue), previously unidentified internal α-N-acetylated peptides (green), internal TNBS nonreacting peptides starting with either Q* or P (red), and an internal Q*/ P peptide containing an apoptosis-specific processing site (yellow). Red crosses delineate protein fragments that were generated by cleavage but further degraded or lost during the sorting procedure.



MSDFDEFERQLNENKQERDKENRHRKRSHSRSRSRDRKRRSRSRDR
RNRDQRSASRDRRRRSKPLTRGAKEEHGGLIRSPRHEKKKKVRKYW
DVPPPGFEHITPMQYKAMQAAGQIPATALLPT*MTPD***GLAVTPTPVP**
VVGSQMTRQARRLYVGNIPFGITEEAMMDFFNAQMRLGGLTQAPGN
PVLAVQINQDKNFAFLEFRSVDETTQAMAFDGIIFQGQSLKIRRPH
DYQPLPGMSENPSVYVPGVVSTVVPDSAHKLFIGGLPNYLNDDQVK
ELLTSFGPLKAFNLVKDSATGLSKGYAFCEYVDINVTDQAIAGLNG
MQLGDKKLLVQRASVGAKNATLVSPPSTINQTPVTLQVPGLMSSQV
QMGGHPTEVLCLMNMVLPEELLDDEEYEEIVEDVRDECSKYGLVKS
IEIPRPVDGVEVPGCGKIFVEFTSVFDCQKAMQGLTGRKFANRVVV
TKYCDPDSYHRRDFW

Figure 2 | Identification of cleavage sites in death substrates. This illustration represents the majority of the cleavages and corresponds to the scenario in **Figure 1b.** The amino-acid sequence of the serine- and arginine-rich splicing factor U2AF 65 kDa subunit is at the bottom. The peptide corresponding to the eight N-terminal residues (underlined) was recovered as an isotopic couple separated by 4 Da (upper left). The new peptide derived from an apoptosis-specific cleavage (bold) was recovered as the ^{16}O-variant (upper right). The tandem MS fragmentation spectrum of this internal peptide (middle) provides data for its unambiguous assignment (the *b*-type fragment ions are italic and the *y*-type are roman; the number in bold corresponds to the mass of the intact peptide ion (singly charged)). The deduced sequence is preceded by a potential caspase cleavage motif shown in italic in the protein sequence.

94%) does not appear to be the result of protein processing (**Fig. 1a**) and is present in highly comparable concentrations in the proteome digests of living and apoptotic cells (**Supplementary Fig. 1** online). Most of these peptides are located at the very N terminus of proteins, and a fraction contains acetylated peptides derived from processing, which was similar in both apoptotic and living cells (for example, the removal of transit peptides upon mitochondrial import, **Supplementary Table 1** online).

Direct identification of protein cleavage sites

The simplest scenario is a specific cleavage in the apoptotic proteome yielding stable products. One could then expect to find the N-terminal peptides of the parent protein in equal concentrations, next to a new internal α-N-acetylated peptide (**Fig. 1b**). We illustrate this scenario for the serine- and arginine-rich splicing factor U2AF 65 kDa subunit (**Fig. 2**). Next to the N-terminal peptide doublet (Ac-SDFDEFER; residues 2–9) in a 1.04 ratio, we identified a second, but now singlet peptide (Ac-GLAVTPTPVPVVGSQMTR; residues 129–146) owing to cleavage in apoptotic cells at the MetThrProAsp consensus caspase site[18]. A list of 93 such directly assigned proteolytic sites located in 71 different proteins from apoptotic cells is available (**Supplementary Table 2** online). Of these cleavages, 58 correspond to caspase consensus sites (**Table 1** and **Supplementary Table 2**). We validated some previously unidentified apoptotic protein processing events by probing the lysates of control and apoptotic cells (in the presence or absence of the caspase inhibitor z-DEVD-FMK[19]) with antibodies specific for three new and one known substrate[20] (PARP) and visualizing protein processing by immunoblotting (**Fig. 3**).

Table 1 | Internally located α-N-acetylated peptides generated after aspartic acid specific cleavage

Protein description	Accession number	Site	Identified peptide	Start	End	Caspase (cleavage yield)
Dynamin 2[a]	P50570	DQVD↓	Ac-TLELSGGAR	353	361	
G-patch domain and KOW motifs	12653257	DSGD↓	Ac-GAGPSPEEKDFLKTVEGR	38	55	
		ALAD↓	Ac-GVVSQAVKELIAESKKSLEER	99	119	
		DRQD↓	Ac-GPAAKSEKAAPR	342	353	
Lysyl-tRNA synthetase[a]	Q15046	VKVD↓	Ac-GSEPKLSKNELKR	13	25	Caspase 3 (65%), 6 (25%) and 8 (45%)
Splicing factor 3B subunit 2[a]	Q13435	TEED↓	Ac-TVSVSKKEKNR	292	302	
Transcription intermediary factor 1-β[a]	Q13263	LSLD↓	Ac-GADSTGVVAKLSPANQR	686	702	
		DGAD↓	Ac-STGVVAKLSPANQR	689	702	Caspase 3 (20%) and 8 (20%)

Five representative members out of the complete list of internally located α-N-acetylated peptides (**Supplementary Table 2**) are shown. Their parent proteins are referred to by description and UniProt database accession number. Information on the amino acid(s) preceding the identified peptides (site), the sequence of the identified peptide and its location within the parent protein (start/end) is given. Ac- denotes the α-N-acetyl group. The right column shows the cleavage yields (in percent) of the corresponding synthetic peptides by recombinant caspases 3, 6, 7 and 8.
[a]Processing events validated in this study by immunoblotting and/or *in vitro* peptide cleavage.

We identified several peptides indicative for protein processing only once (**Supplementary Table 3** online). We validated the deduced cleavage site using synthetic peptides containing these sites as *in vitro* substrates of different recombinant caspases. A list of these validated 'single-hit wonders' is available in **Table 1** and **Supplementary Table 2**.

Notably, we found 35 sites at nonaspartate residues (**Supplementary Table 2**) suggesting caspase cleavage at noncanonical sites (for example, as previously observed for cleavage at Glu10 in the transcription factor Max[21]) and/or additional activation of one or more proteases (see below).

Indirect characterization of protein cleavage

We observed some α-N-acetylated peptides only in the proteome digest of living cells (**Table 2** and **Supplementary Table 4** online). The scenarios depicted in **Figure 1c,d** explain this observation; apoptosis-specific cleavage within the utmost N-terminal peptide renders it unsuited for analysis. Thus, only the corresponding, intact N-terminal peptide in the proteome digest of the living cells can be recovered (**Fig. 1c**). We noticed this scenario (cleavage close to the protein's N terminus) for 12 proteins (**Supplementary Table 4**), and it is illustrated by the observed cleavage pattern of the lysyl-tRNA synthase. For this protein, we identified the acetylated N-terminal peptide spanning residues 2–25 in living cells but detected the peptide indicating processing at Asp12 (amino acids 13–25) only in the proteome digest of apoptotic cells (**Table 1**). The complementary segment 2 to 12 was most likely lost during sample preparation.

We obtained further support for caspase-specific cleavage in the extreme N-terminal regions using an *in vitro* peptide cleavage assay as described above. This way, we delineated the exact cleavage site, the actual caspase(s) involved and the cleavage yield for four extreme N termini that we only detected in lysates of living cells (**Table 2**). Cleavages at Asp8 of the DNAJC7 protein, Asp12 of GCIP-interacting protein p29 isoform 1, Asp12 of lysyl-tRNA synthetase and at Asp8 of Ras GTPase-activating-like protein, were hereby confirmed.

The exclusive presence of peptides in the lysate of living cells could also be explained when the N-terminal protein half has been extensively degraded in apoptotic cells. Such a scenario (**Fig. 1d**) could account for the presence of sorted peptides that do not contain an obvious caspase-specific cleavage site (**Supplementary Table 4**).

We also detected 13 peptides in the proteome digest of living cells that were not acetylated but started with a

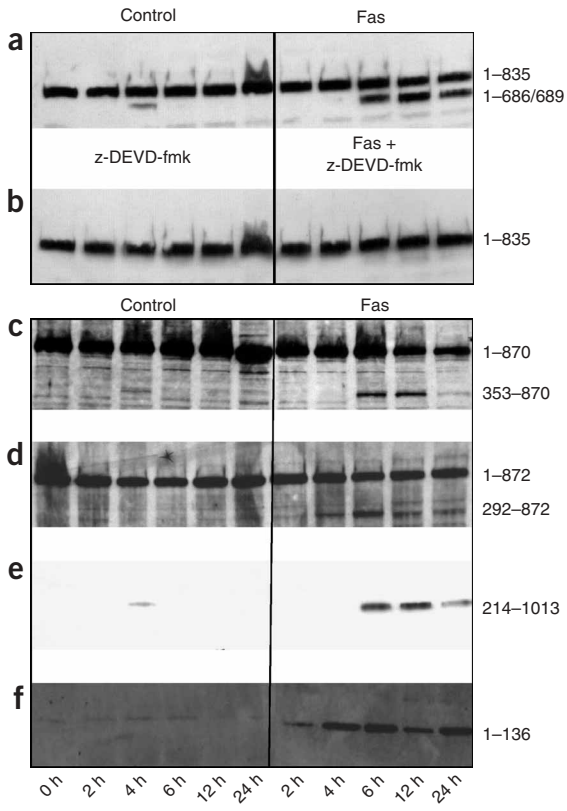


Figure 3 | Validation of protein processing in apoptotic Jurkat cells. (**a**–**f**) Jurkat cells were treated with a Fas antibody for various periods of time after which an aliquot was withdrawn and the proteins separated by SDS-PAGE. As control, cells were identically treated in the same medium but in the absence of a Fas antibody (Control). Antibodies to TIF1β (**a**), dynamin 2 (**c**), subunit 2 of splicing factor 3B (**d**), the caspase-3 fragment of PARP-1 (**e**) and histone H3 (**f**) were used in the immunoblot. TIF1-β detection in presence of a caspase inhibitor is shown in **b**. Note the disappearance of generated fragments in **c** and **d** at later time points suggesting degradation. Positions indicating the boundaries of intact and generated fragments are given in the right margin. They are derived from sequences in databases or from results shown in **Table 1** and **Supplementary Table 2**.

Table 2 | Peptides only present in lysates of living cells

Protein description	Accession number	Identified peptide	Start	End	Site	Caspase (cleavage yield)
N-acetylated peptides						
DnaJ homolog subfamily C member 7[a]	Q99615	Ac-AAAAECDVVMAATEPELLDDQEAKR	2	26	AECD↓	Caspase 8 (40%)
GCIP-interacting protein p29 isoform 1[a]	7661636	Ac-AAIAASEVLVDSAEEGSLAAAAELAAQKR	2	30	VLVD↓	Caspase 3 (40%), 6 (40%) and 8 (55%)
Lysyl-tRNA synthetase[a]	Q15046	Ac-AAVQAAEVKVDGSEPKLSKNELKR	2	25	VKVD↓	Caspase 3 (65%), 6 (25%) and 8 (45%)
Ras GTPase-activating-like protein IQGAP1[a]	P46940	Ac-SAADEVDGLGVAR	2	14	DEVD↓	Caspase 3 (100%), 6 (5%), 7 (100%) and 8 (40%)
TNBS nonreactive peptides						
G-patch domain and KOW motifs	12653257	Q*DGPAAKSEKAAPR	340	353	DRQD↓	

Five representative members out of the complete list of peptides only present in lysates of living cells (**Supplementary Table 4**). Peptides shown here indirectly indicate proteolytic processing. The sequence, location and published information is as in **Table 1**. The underlined residues indicate the processing sites that were validated by either *in vitro* peptide cleavage or by additionally identified peptides. Proteins cleaved at specific positions (Site) are indicated. Q* refers to a PCA derivative. [a]Processing events validated in this study by *in vitro* peptide cleavage.

2,4,6-trinitrobenzenesulfonic acid (TNBS) nonreactive N terminus (**Table 2** and **Supplementary Table 4**). Such peptides are likely to contain a proteolytic site leading to the disappearance of the corresponding peptide in apoptotic cell lysates, even though they were recovered in intact form in lysates of living cells (**Fig. 1e**). An example hereof is the processing of the G-patch domain–and KOW motif–containing T54 protein. For this protein, we identified three acetylated peptides in the apoptotic proteome in accordance with specific cleavage at Asp37, Asp98 and Asp341 (**Table 1**), next to a peptide that was only present in the control proteome, starting with an N-terminal pyrrolidone carboxylic acid (PCA), spanning amino acids 340–353 and thus containing the Asp341 site (**Table 2**). Singlet peptides with PCA or proline could also appear when, after processing of the parent protein, the generated fragments are further degraded under apoptotic conditions (**Fig. 1f**). This could explain the presence of internal peptides that do not have an expected cleavage site such as, for instance, peptide 310–320 of vimentin (**Supplementary Table 4**), which may have lost its apoptotic counterpart owing to cleavage at multiple sites of the parent protein.

The combined information from these two types of sorted peptides unique to living cells now adds at least 28 more indirectly assigned apoptosis-specific cleavage events, eventually leading to the identification of 92 proteins that are specifically degraded during apoptosis (**Supplementary Table 5** online).

Notably, we found 23 peptides solely in the proteome digest of apoptotic cells (**Supplementary Table 6** online). As their corresponding proteins are mainly nuclear, we believe that their appearance might be due to their more efficient extraction upon apoptosis-induced chromatin condensation[22,23]. We also confirmed this by western blot analysis using anti–histone H3 (**Fig. 3f**), revealing specific extraction of intact H3 during lysate preparation of apoptotic cells.

DISCUSSION

We used a gel-free proteomics approach to isolate protein N-terminal peptides. This created a less complex peptide mixture keeping its representative status for the parent proteins. Here we did not focus on generating an exhaustive list of proteins expressed in human Jurkat T lymphocytes, but rather on studying apoptosis-dependent protein processing. As protein processing leads to the generation of new protein N termini, we strongly believe that our approach is well-suited to study global effects of an activated proteolytic degradome. Indeed, not only the protease substrates are distinguished in a complex background, but also, and more importantly, the exact cleavage sites are delineated. This combined information is almost impossible to collect using a single genetic or traditional, gel-based proteomics technique.

There is only one limitation to the N-terminal COFRADIC approach, which is linked to the size of the sorted N-terminal peptide. The distance between the N terminus and the first arginine in the sequence determines this size of the peptide. In the study presented here, the average length of the N-terminal peptides was between 12 and 13 residues, ideal for obtaining good mass spectrometric fragmentation coverage. But when the sorted peptides are either too short or too long, they may become undetectable; the former because they elute in the void volume during chromatography and the latter because they are insoluble. A typical example of the latter is the known processing of PARP-1 at Asp200, which we observed in our samples by immunoblot analysis (**Fig. 3e**) but did not detect by COFRADIC. In fact, in this case, cleavage at Asp200 generated a new N-terminal peptide containing 67 amino acids with 15 ε-N-acetylated lysines. It is likely that this peptide has unusual solubility properties, making its separation by reverse-phase high performance liquid chromatography (RP-HPLC) very difficult. Taken at the level of the total human proteome, and assuming only N-terminal peptides with size between 5 and 40 residues are detectable, the N-terminal COFRADIC approach would theoretically cover 80% of the predicted proteins.

We directly identified 93 cleavage sites, 86 of which (92.5%) had not been noticed previously. We confirmed some of these cleavages by immunoblotting using specific antibodies (**Fig. 3**). In addition, we obtained strong but indirect evidence for apoptosis-specific cleavage at 28 more sites in 21 different proteins. Of these, we confirmed seven and correctly located them using recombinant caspase–mediated cleavage of peptide mimetics (**Tables 1** and **2**, and **Supplementary Tables 2** and **4**). Using this approach, we were able to link specific caspases with particular substrates, like proteasome activator complex subunit-3 as a substrate of caspases 7 and 3.

As expected in the apoptotic Jurkat model, many of the isolated peptides point to caspase-mediated cleavage (**Table 1** and **Supplementary Table 2**); however, 35 processing events occurred C-terminally to nonaspartate residues, mainly involving either basic or bulky hydrophobic amino acids (**Supplementary Table 2**) suggesting apoptosis-dependent activities of at least two other protease families. At this stage it is difficult to state if the

nonaspartate cleavages are the result of a direct activation process or if these are secondary, postcaspase events, for instance, produced by aminopeptidases. For the Acinus protein it is known that an initial caspase-3 cleavage is necessary before a second noncaspase hydrolysis leads to its DNA condensing activity[24]. In our study we observed such 'double heterogeneous' cleavages in several candidate proteins, which could follow a similar activation process. These are, for instance: the nascent polypeptide–associated complex α polypeptide, the polypyrimidine tract–binding protein 1, the SET protein and vimentin (**Supplementary Table 2**). Notably, the set of proteins affected by nonaspartate cleavages includes eight histone proteins, and none of their 14 different processing sites have been reported before. Because of their nature, generally being quite small and highly basic, such proteins may have escaped conventional proteome analysis. Histone H4 is processed such that only a few amino acids are removed (cleavage sites at Arg3, Gly4 and Lys5). Cleavage here is noticed at nonaspartate residues, yet a remarkable number of other proteins appears to be specifically cleaved at aspartate close to their N termini (for example, RING-box protein 1 (Asp8), nucleolar protein family 1, member 2 (Asp9), polypyrimidine tract–binding protein 1 (Asp3 and Asp8) and LSm3 (Asp5)). Such processing is difficult to detect by one-dimensional gel shotgun proteomics approaches[25] because the extent of processing is generally not sufficient to evoke a substantial shift in the apparent molecular weight.

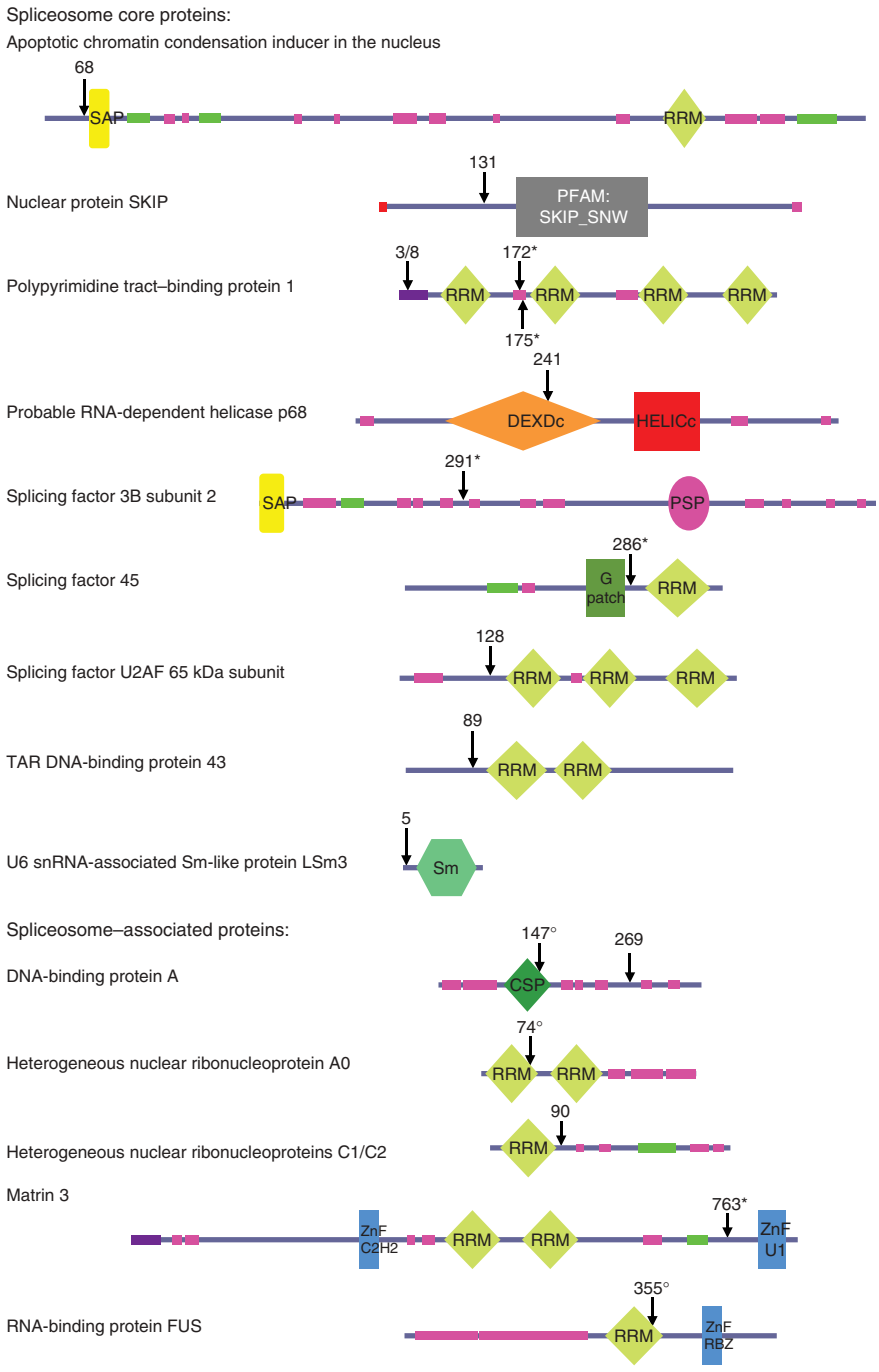


Figure 4 | Location of the apoptotic-specific cleavage sites within the predicted domain organizations of affected spliceosomal proteins. Spliceosomal proteins are subdivided into core and associated proteins. Proteins are shown as blue bars. The domains or motifs are predicted by the SMART algorithm (http://smart.embl-heidelberg.de/) and pictographically presented, and their PFAM nomenclature is indicated (http://www.sanger.ac.uk/Software/Pfam/). Unnamed stretches of amino acids are predicted signal peptides (purple), coiled-coil regions (green) and regions of low complexity (magenta). The cleavage site is indicated with a vertical arrow and its position within the affected protein is given. Cleavages assumed to separate functional entities (*) or to destroy these entities (°) are indicated.

The identified protease substrates are involved in all main biological processes, suggesting that every cellular activity is affected during apoptosis. In particular, we found an unusually large number of DNA- and RNA-binding proteins. Especially noteworthy is that we identified 14 proteins known to be associated with the spliceosome[26–29] and, except for two cleavage sites in the polypyrimidine tract–binding protein 1, the actual sites of protein processing had not been determined previously. We used SMART[30] (http://smart.embl-heidelberg.de/) to analyze how proteolytic processing may interfere with the function of these spliceosomal proteins (**Fig. 4**). Some proteins are predicted to be cleaved within one of their functional entities or processed such that these become separated (**Fig. 4**). The former may lead to loss of function, whereas the latter could suggest differential regulation or localization.

Given that alternative splicing has been reported as an important aspect of apoptosis[31], our observations hint to an (at least partial) interference of RNA splicing by apoptotic proteases. This may lead to the synthesis of protein variants important in the execution of the apoptotic process or in signaling this event to neighboring cells, a question that can be addressed in future

experiments. Similar SMART analyses of all other processed proteins are available in **Supplementary Figure 2** online.

As protein processing has been recognized to be a key regulator of many processes (for example, protein translocation, complement activation, enzyme activation, viral maturation) and perturbations of the protease-antiprotease balance have been considered as the major cause of major diseases (such as sepsis, chronic obstruction pulmonary disease, Alzheimer disease, cancer metastasis and others.) the differential N-terminal COFRADIC procedure offers an unique tool for routine screening of substrate repertoires of natural or pathological protein processing events.

METHODS

Sorting for N-terminal peptides. The N-terminal peptide sorting procedure consists of three essential steps. First we alkylated proteins present in the lysate and then blocked them by acetylation of free amino groups followed by trypsin cleavage. We then fractionated the peptide mixture by RP-HPLC (primary run). We treated peptides in each fraction with TNBS, converting all peptides with a free N terminus into hydrophobic trinitrophenyl derivatives. Peptides with a blocked α-N-acetylated or PCA terminus and peptides with a nonreactive N terminus (proline) will not be modified. During a replicate secondary run, the latter elute in the same time interval, whereas the internal, now trinitrophenyl peptides, shift to later elution times. When this procedure is repeated for each of the primary fractions, the complete mixture is finally sorted into two peptide sets. The unaltered peptides that do not shift during the two consecutive runs can be easily collected in their previously recorded time interval and are retained for mass spectrometric identification. These are mainly α-N-acetylated peptides derived from the protein N termini, or internal peptides starting with PCA or proline (**Fig. 1**).

The quantitative differential aspect of the procedure is based on postcleavage, trypsin-catalyzed C-terminal ^{18}O exchange, creating a mass difference of 4 Da compared to the untreated peptide mixture[17]. Tryptic peptides from apoptotic lysates were generally left with the normal oxygen isotope, whereas those from living cells were tagged with ^{18}O. In one experiment we swapped the oxygen labels to verify the general quantitative aspect of the ^{18}O exchange procedure.

Additional methods. Experimental details of the COFRADIC procedure, cell culture, induction of apoptosis and associated biochemical experiments are available in **Supplementary Methods** online.

Accession codes. BIND identifiers (http://bind.ca): 316861, 316862, 316863, 316864, 316865, 316866, 316867, 316868, 316869, 316870, 316871, 316872, 316873, 316874, 316875, 316876, 316877, 316878, 316879, 316880, 316881, 316882, 316883, 316884, 316885, 316886, 316887 and 316888.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

The authors thank P. Vandenabeele for his critical reading of the manuscript, for the suggestions made and for providing the recombinant caspases. K.G. is a postdoctoral fellow and L.M. is a research assistant of the Fund for Scientific Research-Flanders (Belgium; F.W.O.-Vlaanderen). The project was supported by research grants from the Fund for Scientific Research-Flanders (Belgium; project number G.0008.03), the Inter University Attraction Poles (IUAP, project number P5/05), the GBOU research initiative (project number 20204) of the Flanders Institute of Science and Technology (IWT) and the European Union Interaction Proteome (6th Framework Program).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
2. Dewachter, I. & Van Leuven, F. Secretases as targets for the treatment of Alzheimer's disease: the prospects. *Lancet Neurol.* **1**, 409–416 (2002).
3. Berdowska, I. Cysteine proteases as disease markers. *Clin. Chim. Acta* **342**, 41–69 (2004).
4. Docherty, A.J., Crabbe, T., O'Connell, J.P. & Groom, C.R. Proteases as drug targets. *Biochem. Soc. Symp.* **70**, 147–161 (2003).
5. Lamkanfi, M., Declercq, W., Kalai, M., Saelens, X. & Vandenabeele, P. Alice in caspase land. A phylogenetic analysis of caspases from worm to man. *Cell Death Differ.* **9**, 358–361 (2002).
6. Fischer, U., Janicke, R.U. & Schulze-Osthoff, K. Many cuts to ruin: a comprehensive update of caspase substrates. *Cell Death Differ.* **10**, 76–100 (2003).
7. Cryns, V.L. *et al.* Specific proteolysis of the kinase protein kinase C-related kinase 2 by caspase-3 during apoptosis. Identification by a novel, small pool expression cloning strategy. *J. Biol. Chem.* **272**, 29449–29453 (1997).
8. Kamada, S. *et al.* A cloning method for caspase substrates that uses the yeast two-hybrid system: cloning of the antiapoptotic gene gelsolin. *Proc. Natl. Acad. Sci. USA* **95**, 8532–8537 (1998).
9. Kuida, K. *et al.* Altered cytokine export and apoptosis in mice deficient in interleukin-1β converting enzyme. *Science* **267**, 2000–2003 (1995).
10. Li, P. *et al.* Mice deficient in IL-1β-converting enzyme are defective in production of mature IL-1β and resistant to endotoxic shock. *Cell* **80**, 401–411 (1995).
11. Gerner, C. *et al.* The Fas-induced apoptosis analyzed by high throughput proteome analysis. *J. Biol. Chem.* **275**, 39018–39026 (2000).
12. Thiede, B., Dimmler, C., Siejak, F. & Rudel, T. Predominant identification of RNA-binding proteins in Fas-induced apoptosis by proteome analysis. *J. Biol. Chem.* **276**, 26044–26050 (2001).
13. Gerner, C. *et al.* Proteome analysis of nuclear matrix proteins during apoptotic chromatin condensation. *Cell Death Differ.* **9**, 671–681 (2002).
14. Thiede, B., Siejak, F., Dimmler, C. & Rudel, T. Prediction of translocation and cleavage of heterogeneous ribonuclear proteins and Rho guanine nucleotide dissociation inhibitor 2 during apoptosis by subcellular proteome analysis. *Proteomics* **2**, 996–1006 (2002).
15. Wajant, H. The Fas signalling pathway: more than a paradigm. *Science* **296**, 1635–1636 (2002).
16. Gevaert, K. *et al.* Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat. Biotechnol.* **21**, 566–569 (2003).
17. Staes, A. *et al.* Global differential non-gel proteomics by quantitative and stable labelling of tryptic peptides with oxygen-18. *J. Proteome. Res.* **3**, 786–791 (2004).
18. Thornberry, N.A. *et al.* A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. *J. Biol. Chem.* **272**, 17907–17911 (1997).
19. Garcia-Calvo, M. *et al.* Inhibition of human caspases by peptide-based and macromolecular inhibitors. *J. Biol. Chem.* **273**, 32608–32613 (1998).
20. Nicholson, D.W. *et al.* Identification and inhibition of the ICE/CED-3 protease necessary for mammalian apoptosis. *Nature* **376**, 37–43 (1995).
21. Krippner-Heidenreich, A. *et al.* Targeting of the transcription factor Max during apoptosis: phosphorylation-regulated cleavage by caspase-5 at an unusual glutamic acid residue in position P1. *Biochem. J.* **358**, 705–715 (2001).
22. Wu, D. *et al.* Apoptotic release of histones from nucleosomes. *J. Biol. Chem.* **277**, 12001–12008 (2002).
23. Galande, S., Dickinson, L.A., Mian, I.S., Sikorska, M. & Kohwi-Shigematsu, T. SATB1 cleavage by caspase 6 disrupts PDZ domain-mediated dimerization, causing detachment from chromatin early in T-cell apoptosis. *Mol. Cell. Biol.* **21**, 5591–5604 (2001).
24. Sahara, S. *et al.* Acinus is a caspase-3–activated protein required for apoptotic chromatin condensation. *Nature* **401**, 168–173 (1999).

25. Thiede, B., Treumann, A., Kretschmer, A., Sohlke, J. & Rudel, T. Shotgun proteome analysis of protein cleavage in apoptotic cells. *Proteomics* **5**, 2123–2130 (2005).
26. Buratti, E. & Baralle, F.E. Characterization and functional implications of the RNA binding properties of nuclear factor TDP-43, a novel splicing regulator of CFTR exon 9. *J. Biol. Chem.* **276**, 36337–36343 (2001).
27. Lallena, M.J., Chalmers, K.J., Llamazares, S., Lamond, A.I. & Valcarcel, J. Splicing regulation at the second catalytic step by Sex-lethal involves 3′ splice site recognition by SPF45. *Cell* **109**, 285–296 (2002).
28. Zhou, Z., Licklider, L.J., Gygi, S.P. & Reed, R. Comprehensive proteomic analysis of the human spliceosome. *Nature* **419**, 182–185 (2002).
29. Schwerk, C. *et al.* ASAP, a novel protein complex involved in RNA processing and apoptosis. *Mol. Cell. Biol.* **23**, 2981–2990 (2003).
30. Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. SMART, a simple modular architecture research tool: identification of signalling domains. *Proc. Natl. Acad. Sci. USA* **95**, 5857–5864 (1998).
31. Jiang, Z.H. & Wu, J.Y. Alternative splicing and programmed cell death. *Proc. Soc. Exp. Biol. Med.* **220**, 64–72 (1999).

2.2. ms_lims: channeling the flood of data

"First, the tools were primitive in the features they offered. Second, the features, when available, often had built-in limitations that tended to break when projects started to get complicated. Third, support from proprietary vendors was terrible; unless you were in the process of buying lots of hardware or renewing a large software site license, and could use the power of the purse to your advantage, you were out of luck when you ran into one of these built-in limitations. And finally, every vendor implemented their own proprietary extensions, so that when you did use the meager features of one platform, you became, imperceptibly at first, then more obviously later, inextricably tied to that platform. All in all, it was quite clear that whatever the merits of free market economics, they were not at work in the software marketplace. The extent to which the proprietary software model was a broken model made the study of that model extremely valuable indeed."

- Michael Tiemann, *Open Sources: Voices from the Open Source Revolution*

In order to fulfill the software requirements listed in section 1.4, any single lab can choose from two options. The first of these is to buy a Laboratory Information Management System (LIMS) from the marketplace and the second is to develop a system in-house.

The commercial options fall into two broad categories: 'generic' solutions and mass spectrometer vendor solutions.

The generic solutions are first and foremost prohibitively expensive. These systems are usually simply re-packagings of applications developed for the pharmaceutical industry for drug discovery and development. As such these systems have been designed around rigid procedures with rigorously defined user roles and access privileges combined with extensive auditing on every data manipulation operation. These features are of course essential when filing for drug approval, but can hamper purely scientific efforts because of their enormous hardware requirements, slow maintenance and update cycles and lack of transparency. Additionally, several core modules (e.g. integration of a specific data producing instrument) are often missing and can only be provided on a per-user basis by highly expensive consultants who will usually write a one-off solution that can not accommodate the shifting requirements that typify basic research.

The second type of commercial system is represented by the information management tools developed by the instrument vendors. Their systems are typically user-friendly, fully integrated management systems that automate and structure common tasks well enough. The caveat is that integration of instruments by other vendors is typically wholly absent. Although efforts are sometimes made to compensate for this, many of the most useful features are then inaccessible for the other instruments and no single provider can guarantee support for future instrument purchases if these instruments are from a different vendor[26]. Additionally, since the core software development business of these companies is related to instrument control software[27], backwards compatibility is not (yet) an important issue for them and can almost never be guaranteed.

[26] It is certainly no exaggeration that many vendors do not even want to commit themselves to guaranteeing support for their own future models!

[27] Which changes version with each new model, nearly always introducing several new features.

Freely available solutions were simply non-existent when ms_lims was first created, although in the few last years, several systems have been published [Field 2002, Kristensen 2004, Rauch 2006], or have been made available without having been published as a whole (http://www.sbeams.org, http://www.thegpm.org). It is indicative of the poor availability of usable commercial systems that development on these published systems started at around the same time in different laboratories across the globe to fill the urgent need for automated data storage and analysis.
In the proteomics group in Ghent, the ms_lims software was developed to tackle the data management problems. It builds on a centralized relational database for structured storage of the data and provides relevant applications for data storage, data processing and data analysis and presentation. The relational schema will be discussed first, as well as two auxiliary applications that allow for a more efficient development and deployment of the ms_lims data access layer. An overview of the ms_lims applications will be given next, broken down by the three categories explained above. Finally, published applications of the ms_lims system will be highlighted.

2.2.1. The database schema

"Show me your functions, and I will be confused. Show me your data, and your functions will be obvious."

- Frederick P. Brooks, *The Mythical Man-Month*

The ms_lims relational database schema is depicted in figure 34.
It is of note that, of the fourteen tables, only five[28] are strictly related to identifications. This leaves nine tables, of which two[29] support supplemental storage of binary files and one[30] is an application specific, descriptive class-loading table. The remaining six[31] tables are dedicated to source data capture, tracing and integration.
The central table is the 'project' table. All other tables are connected to this table, except for the 'projectanalyzertool' table.
An important feature of most tables[32] is the presence of rudimentary auditing columns: '*username*', '*creationdate*' and '*modificationdate*'. These columns allow the tracing through time of the user that triggered the insertion of the row, the time and date on which this insert was accomplished and the time and date of the last modification to the row.
The occurrence of a BLOB[33] or LONGBLOB column type highlights where binary files can be stored in the database. Each binary column in the ms_lims database contains zipped or GZIPped data[34] to minimize storage space and table size. Conversion from

[28] These are 'identification', 'datfile', 'status', 'phosphorylation' and the indirection table 'id_to_phospho'.
[29] 'binfile' and 'filedescriptor'; examples of possible supplementary binary files include: an Excel sheet, a Word document, a gel image (e.g. in JPEG format) and a set of files comprising the output of a third-party analysis tool.
[30] 'projectanalyzertool' which is queried exclusively by the ProjectAnalyzer application.
[31] 'user', 'cofradic', 'project', 'instrument', 'lcrun' and 'spectrumfile'.
[32] The only exceptions are the three phosphorylation related tables.
[33] BLOB is an acronym for 'Binary Large OBject'
[34] Zip compression is used for entities comprising multiple entries, such as folders. GZIP is used for the storage of single streams such as individual files.

compressed to uncompressed data is handled automatically by the relevant ms_lims applications but, thanks to the usage of common compression standards, third-party consumers can readily decompress the information after retrieval as well.

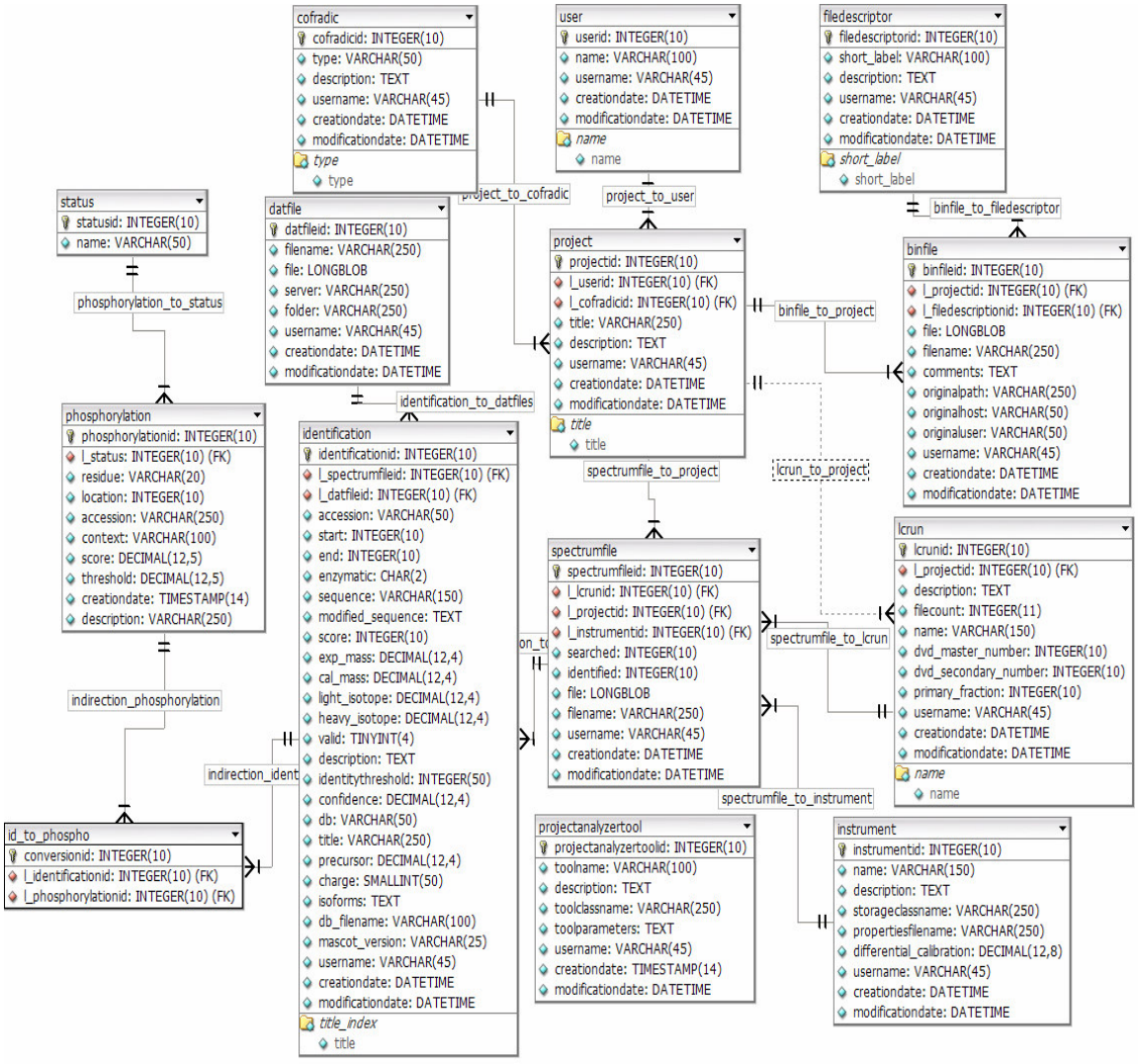


Figure 34: ms_lims version 5.0.x relational database schema, 'crows-feet' notation.

2.2.2. Building on solid ground: data access code generator

Data access is the most vital part of any data management software system. Indeed, if the data access layer does not retrieve (or even worse: store) the data correctly, the usability of the entire system becomes zero. The data access layer must therefore consist of proven, reliable code that at the same time performs reasonably well. Another desired characteristic of the data access layer is the ability to abstract its primary functions. Finally, it is also important to provide quick and efficient updates of the data access layer whenever the underlying data store changes.

A LIMS system for a proteomics research group makes no exceptions to these basic requirements. The dynamic research setting even emphasizes the importance of the

adaptability of the data access layer, as the underlying database schema is prone to change frequently and often dramatically as the techniques applied or the scientific topics researched evolve.
The most efficient way to create a highly adaptable yet reliable and efficient data access layer is not to write one at all. Instead, the task of procuring the code should be delegated to a software program that outputs reliable, proven code in perfectly reproducible ways and at speeds unrivaled by any other process.
We have built such a general-purpose table-based code generator which outputs code that fits into a generic framework for data access, consisting of four interfaces each of which defines one of the four basic data operations: *Persistable*, *Updateable*, *Retrievable* and *Deleteable*. The generated data access components take care of their own data access operations and require only a standard Java Connection object on which their operations should be performed, effectively abstracting the data layer from the business logic layer. These data access components are also self-aware with regards to changes to their internal data. This functionality can be described in a UML state diagram (figure 35) and a single component thus represents a simplistic finite state machine.

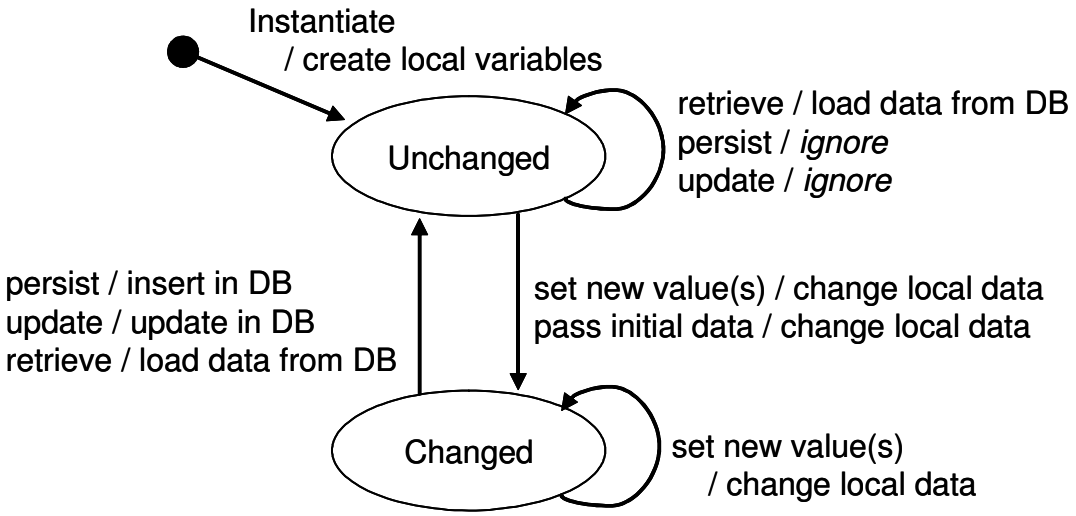


Figure 35: UML state diagram for a generated data accessor component.

By extending the generated code to add custom extensions, it becomes possible to regenerate the data access component whenever the underlying database schema changes without suffering from downstream effects[35].
The data access generator has proven its usefulness repeatedly over the years as many changes have occurred to the ms_lims database schema since its inception.

[35] Typically there are some effects such as changed column names that can require a few minor modifications to the extending class. Major changes (e.g. the addition of one or more columns) usually also require updates to the user interface to display these new data fields, yet without actually making any of these changes, the software will still function, albeit sub-optimally as it does not yet show all data available.

2.2.3. Building on shifting ground: database migration tool

"All successful software gets changed. Two processes are at work. First, as a software product is found to be useful, people try it in new cases at the edge of or beyond the original domain. The pressures for extended function come chiefly from users who like the basic function and invent new uses for it."

- Frederick P. Brooks, *The Mythical Man-Month*

Any system that has been sufficiently long in production is almost certain to undergo revisions to its functionalities and underlying data schema. As changes to the latter must be retroactively applied to the data already in the system, these updates can be cumbersome, time-consuming and potentially dangerous[36].

In practice the update can require complex rearrangements of existing data. It is also often the case that temporary changes to the database metadata[37] can significantly increase the speed of converting an existing database into a new format. This leads to multi-step upgrade processes that typically require a set of Structured Query Language (SQL) statements as well as some more complex steps that need to call on small computer programs to manage intricate data transitions. All these steps need to be executed in a strict order, as nearly all of the steps require certain modifications to have been performed by one or more of the preceding steps.

In order to allow fully automatic and monitorable updates of any kind of relational database, the DBTransferTool program was developed. This program reads a structured text file called Conversion Definition File ('.cdf' file) which lists all necessary steps in order. These steps can be of SQL type, resulting in the program executing the statement against the specified database, or of programmatic type, in which case the value of the step should be the fully qualified class name of an implementation of the *DBConverterStep* interface. This interface defines a single method, performConversionStep(Connection aConn) which will be executed by the DBTransferTool.

Execution time as well as possible errors are tracked for each step by the transfer tool, allowing profiling of the steps when applied to (typically smaller) test databases. This feature makes it relatively easy to spot bottle-neck statements, which can then be analyzed and sped up by specifically tweaking them.

Finally, the most important advantage of the tool is that the complete update procedure of one database version to another database version is contained and documented in a single text file. Knowledgeable users can thus directly read all steps[38], understand the changes applied and potentially make modifications where appropriate if they have altered their local version of the database.

---

[36] It therefore deserves recommendation to always provide verified backups before upgrading databases to a new schema.

[37] Examples of such database metadata are column types, indexes and constraints.

[38] Note that direct reading of the programmatic code requires either the source code or reverse engineering of the actual Java binary that represents this step. Since ms_lims is made available under an open source license and the code is freely available, this is not a problem here.

2.2.4. Applications for data entry

A database can only be put to use once it contains data. Data entry is often a crucial step as data already in the database tends to be trusted. It is also the step where different data sources and formats are integrated into a single common schema. Interestingly, apart from the clear advantages this integration effort yields, there is an important caveat: implicit metadata connected to the use of certain formats is lost[39]. It is therefore also the stage where this implicit metadata is converted into a structured description of the origins of the data.

In developing applications for data entry it is a golden rule to automate as much of the processing as possible. User input should be strictly limited to the necessary, user-specific knowledge. This way, the user can maximize her/his attention span and minimize the introduction of human error[40]. After the operation completes, the user should also be presented with (preferentially numerical) feedback on the proceedings, allowing a quick verification step to take place. It is evident that the group of users at large benefits from a uniform interface as the amount of training required to make the transition to a new instrument becomes negligible.

In ms_lims the main data entry application is SpectrumStorageGUI[41]. The application itself is simply a graphical interface on top of a frameworked set of data loaders defined by the *SpectrumStorageEngine* interface. This framework, illustrated in the UML diagram[42] in figure 36, is built around two basic functionalities: locating LC runs and storing these in the database.

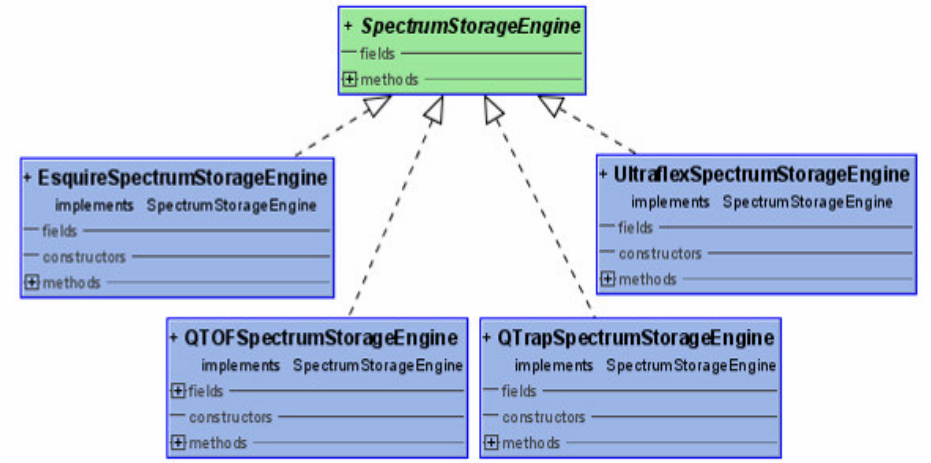


Figure 36: UML diagram of the SpectrumStorageEngine framework in ms_lims.

[39] For example: the use of '.pkl' files hints at a Micromass/Waters instrument while '.xml' files can only be derived from the Bruker Ultraflex.

[40] Human error (as opposed to machine error) tends to be non-systematic and extremely difficult to trace.

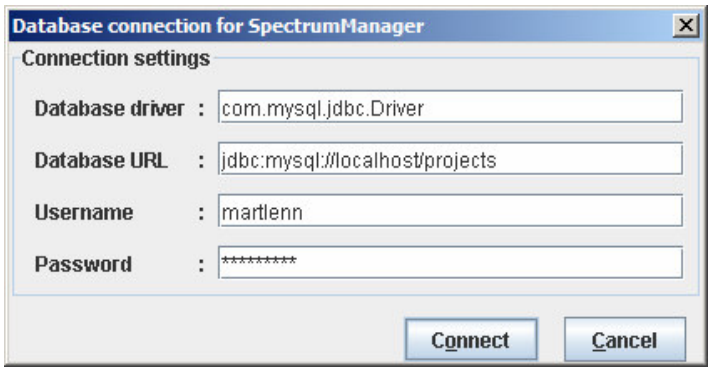
[41] There are other data entry applications used for assigning backup medium numbers and primary LC run fraction numbers, but these are trivial in design and are therefore not discussed in detail.

[42] This UML diagram is not a conceptual model; it is automatically generated on the basis of the actual ms_lims code itself.

The corresponding methods on the *SpectrumStorageEngine* interface are aptly named findAllLCRunsFromFileSystem(…) and loadAndStoreSpectrumFiles(…)[43].
It is worthwhile to walk through usage of the SpectrumStorageGUI application in full, as it is representative of most ms_lims applications.
The first step in any application is managing the connection to the database. This is achieved in each application through the same, shared component: *ConnectionDialog*. A screenshot of this login procedure can be seen in figure 37(a). Upon database connection, the user is informed of this successful login (figure 37(b)) and is subsequently presented with a selection dialog where all available instruments are displayed (figure 37(c)).
The contents of the instrument selection dialog screen are collected from the ms_lims database[44]. The database contains short names (shown in the dropdown box) as well as lengthier descriptions (shown in blue italics above the buttons). Invisible on the interface, but present in memory are the fully qualified class names for the relevant *SpectrumStorageEngine* implementation. This system is highly reminiscent of the dynamic detection of database loaders for DBToolkit described in section 2.1.3, with the main difference being the use of the database for descriptive data storage instead of the text files used by DBToolkit.
Upon selection of the desired instrument, the SpectrumStorageGUI application attempts to dynamically load the corresponding *SpectrumStorageEngine* implementation. When successful, the findAllLCRunsFromFileSystem(…) method is invoked which returns a list of LC runs that are candidates for being moved to the database. During the search for LC runs on the file system, a progress bar keeps the user informed of the proceedings. Another silent process running in the background loads all projects currently in the database.

[43] Note that, for the sake of brevity, the method parameters are omitted.
[44] The table that provides this information is the 'instrument' table.

a)



b)

c)

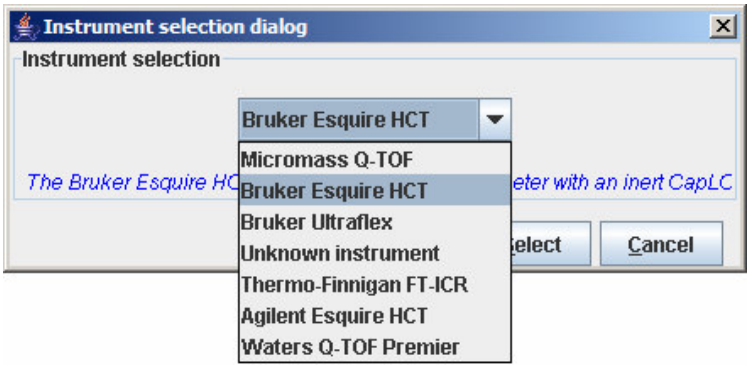


Figure 37: Screenshots of the start-up process of the SpectrumStorageGUI application.

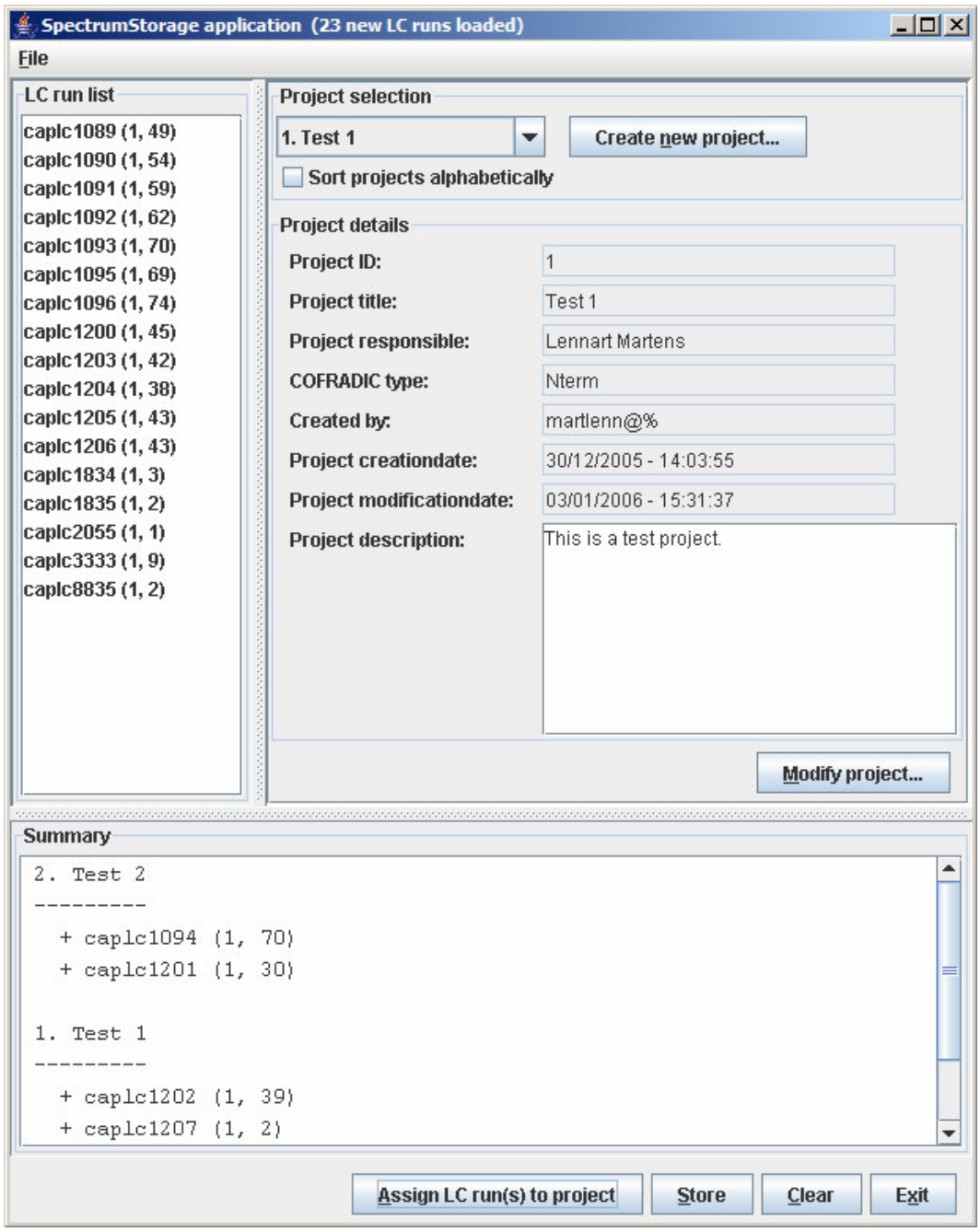
(a) Shared database connection dialog. Database driver and the URL are defined in per-installation text files so the user need not concern herself with them.

(b) Example of informative user feedback: connection has been established and the connected database is repeated for verification purposes.

(c) Instrument selection dialog. Note that there are more instruments in the list than are listed in section 1.1.4. This is due to the fact that ms_lims has been installed at several locations outside of Ghent, however most supported instruments are present for testing purposes in the development database. Note also the 'Unknown instrument', which is a placeholder for testing purposes.

Upon completion of these operations the user is presented with the main application screen, which is shown in figure 38 for two separate instrument selections.

a)



b)

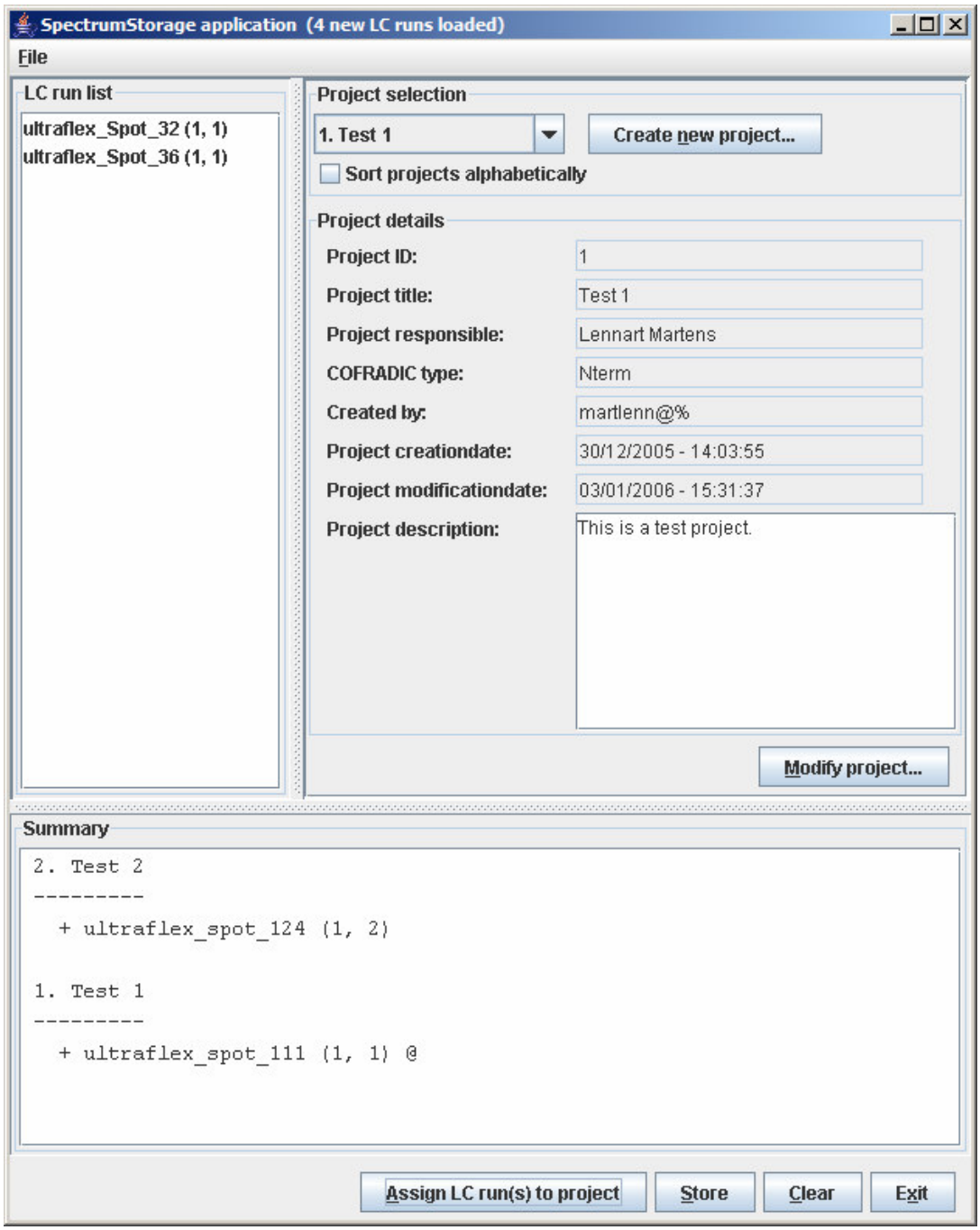


Figure 38: SpectrumStorageGUI main screen with some LC runs already assigned to projects for (a) the Micromass Q-TOF and (b) the Bruker Ultraflex instruments.

It is clear from this figure that the only difference for the application between these two instruments lies in the name of the LC runs. SpectrumStorageGUI thus qualifies for the requirement that the interface should be consistent across multiple instruments. It also automates every task save the assignment of LC runs to projects, which only the machine

operator knows. The possibility for human error is thus reduced. Finally, detailed user feedback is pervasive and unobtrusive across the screen. First and foremost, the title bar indicates the number of LC runs found. The operator can directly compare this number with the expected number (i.e. the number of runs the machine was instructed to perform) and spot any incongruities. All known details for the currently selected project are clearly shown in a prominent location on screen so as to minimize erroneous project assignment. Projects in the dropdown can be sorted by descending project ID (equivalent to reverse chronological sorting) or alphabetical sorting on project title to minimize search time for the desired project. During the assignment process an overview is continuously updated at the bottom, showing all projects that carry assigned LC runs. Each LC run is annotated by a pair of numbers between brackets. These numbers are the total number of child LC runs for a specific run[45] and the number of individual peak lists contained in the LC run. Finally, an LC run can also be affixed with '@'. This implies addition of a user-defined free-text comment to that LC run[46].
When the user is satisfied with the assignments made and wishes to store these in the database, she/he simply needs to click the 'Store' button. The application will show a progress bar, keeping the user informed of the proceedings while converting the instrument specific formatting to ms_lims standard formatting[47] and storing the relevant data in the database[48]. Completion of this phase is concluded by informing the user of the total amount of data stored as shown in figure 39.

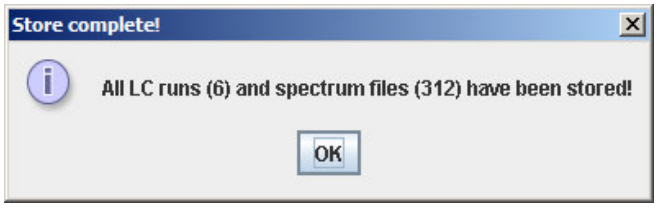


Figure 39: User feedback upon completion of data storage.

One final aspect to discuss with respect to SpectrumStorageGUI concerns its flexibility. Since the data loaders are frameworked and need only implement two methods (one to find all LC runs on the system and the other to store the data[49]) in order to be fully compatible with the SpectrumStorageGUI application, and given the fact that these classes are dynamically loaded from descriptive information in the database, it becomes very easy to add new instruments to the pool[50].

---

[45] Note that this number is always one here because of good data management practice in the settings for the processing software on the instruments.
[46] Double-clicking on an LC run in the list view to the right of the screen will open a dialog that allows free text entry or editing for each LC run.
[47] The current standard format for peak lists in ms_lims is the Mascot Generic Format ('.mgf' files) due to easy integration with the Mascot search engine, the high level of annotation it can carry and the relative ease with which both humans and machines can read it.
[48] The 'lcrun' table will contain the LC runs, the 'spectrumfile' table the (GZIPped) peaklists and the links to the correct instrument in the 'instrument' table.
[49] Remember that most of the data access in ms_lims happens through generated accessor objects, reducing the burden of data storage to simply performing the right calls at the right times.
[50] From experience, the typical amount of work for a knowledgeable developer is half a day, including testing and release of the new code version.

2.2.5. Applications for data processing

As was explained in section 1.4.2 the main concern for applications that automate data processing in a proteomics laboratory is integration with third-party software. For the group in Ghent, this software is Mascot.

When processing spectra into identifications there are two distinct steps that need to be automated. The first is submission of a selection of the spectra in the system to the search engine, the second the retrieval, display and storage of the resulting identifications. Interestingly, the default Mascot software suite already provides a tool that takes care of submitting spectra to the Mascot server called 'Mascot Daemon'. It is a small Visual Basic application that presents a graphical user interface for submissions and builds on an underlying relational database (default database engine is Microsoft Access) for storage and processing of the results. On top of this, query results are also automatically retrieved and pointers to the original result file are stored in the database. Since the functionality in Mascot Daemon is quite extensive, allowing for instance the re-submission of non-identified spectra to other queries (e.g. containing different search parameters or searching a different database) in concatenated chains, it was decided to build ms_lims around this existing tool rather than trying to duplicate this functionality in an in-house designed application.

In order to complement Mascot Daemon, only three things needed to be done: (1) allow selection of the spectra that will be queried from the ms_lims database, (2) merge these selected spectra in one or more "mergefiles" and (3) read and interpret the raw Mascot results after searching has been completed and store these in the database.

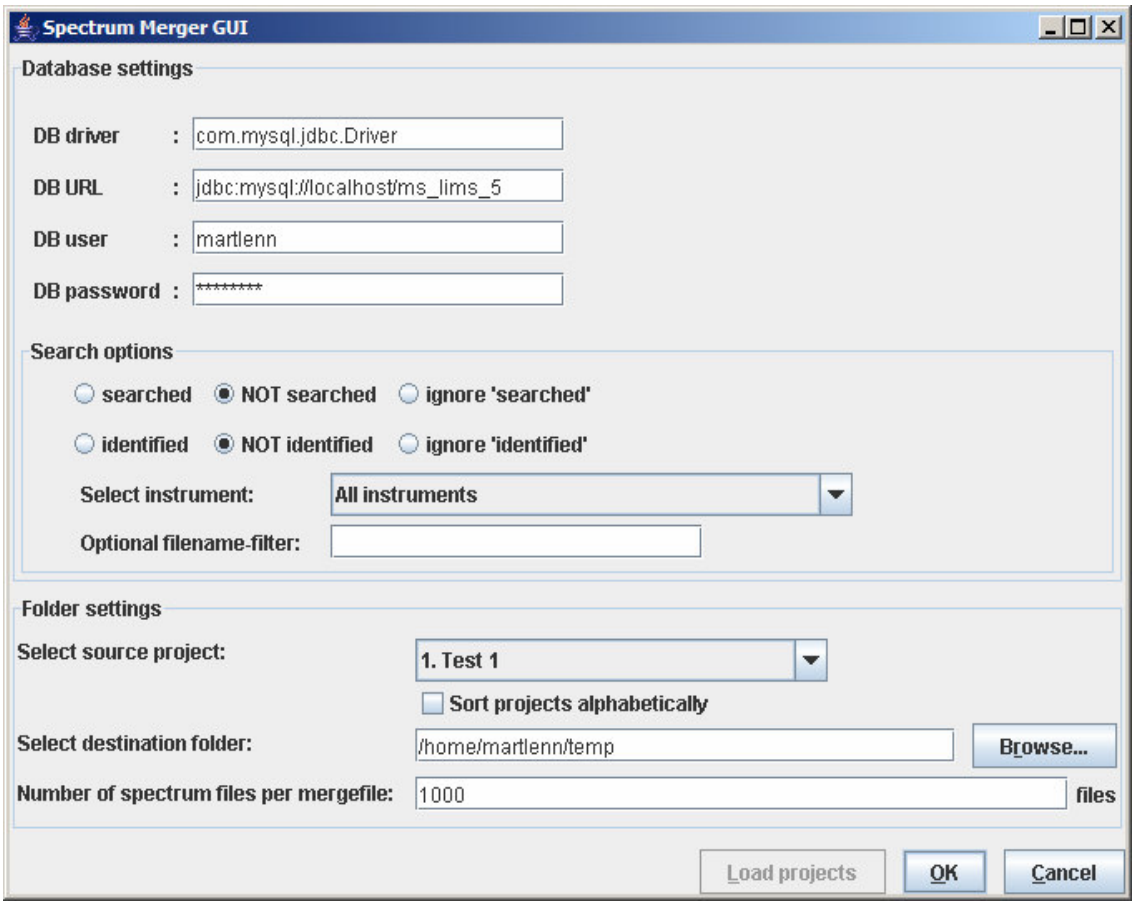
The first and second functionality has been combined in the MergerGUI application, a screenshot of which is shown in figure 40(a). The top panel handles initial database connection, the panel in the middle contains search options and the bottom panel allows project selection and output options.

The middle panel allows the selection of spectra based upon their status (whether or not they have been searched or identified), instrument of origin and, optionally, the spectrum filename. Note that the output panel allows the specification of the number of spectra to merge per mergefile. The creation of these mergefiles carries two important advantages: firstly, the resulting output will be more manageable as there is a thousand fold reduction in the number of files output[51] and secondly, the Mascot searches will become much more efficient. This latter effect has to do with the way Mascot searches spectra. Each file it is presented with causes the execution of a search through a sequence database, regardless of the number of spectra in that file. This means that searching 1000 separate peak lists will result in 1000 independent passes through the same sequence database whereas one mergefile containing 1000 peak lists will result in one pass through the sequence database file[52]. Since the amount of time spent on I/O operations while reading a sequence database constitutes a substantial amount of the total analysis time, the speed gains for merging peak lists in compound mergefiles are very substantial.

[51] Provided of course that 1000 spectra are merged per mergefile.

[52] It is of note that the very latest version of Mascot Daemon (2.1.03, released on the 3rd of November 2005) now has this merging feature built-in.

a)



b)

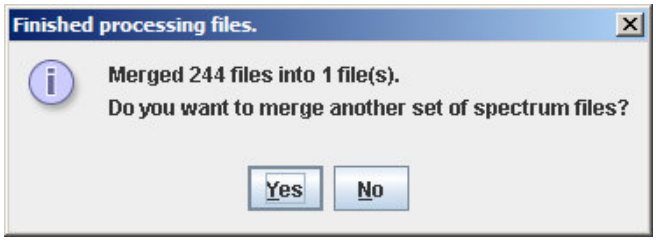
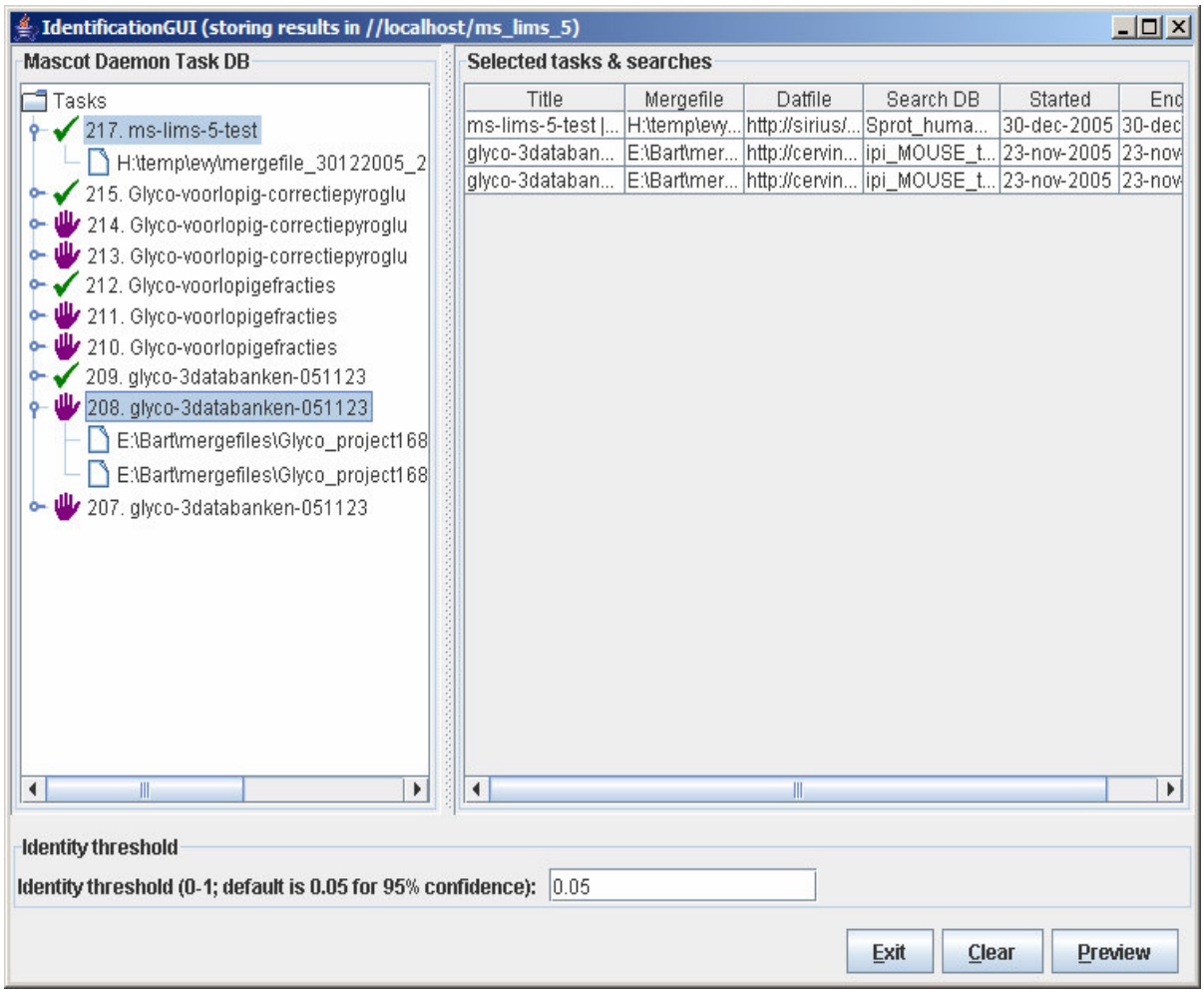


Figure 40: (a) MergerGUI application screenshot.
(b) User feedback on completion of spectrum merging.

The user is presented with an informative progress bar and when the selection, merging and output are complete, a dialog is presented with a numerical report of the actions performed (figure 40(b)).
The third required functionality to complement Mascot Daemon concerns the retrieval, parsing, presentation and storage of the raw Mascot results. These actions are comprised in the IdentificationGUI application, the main screen of which is shown in figure 41(a).

a)



b)

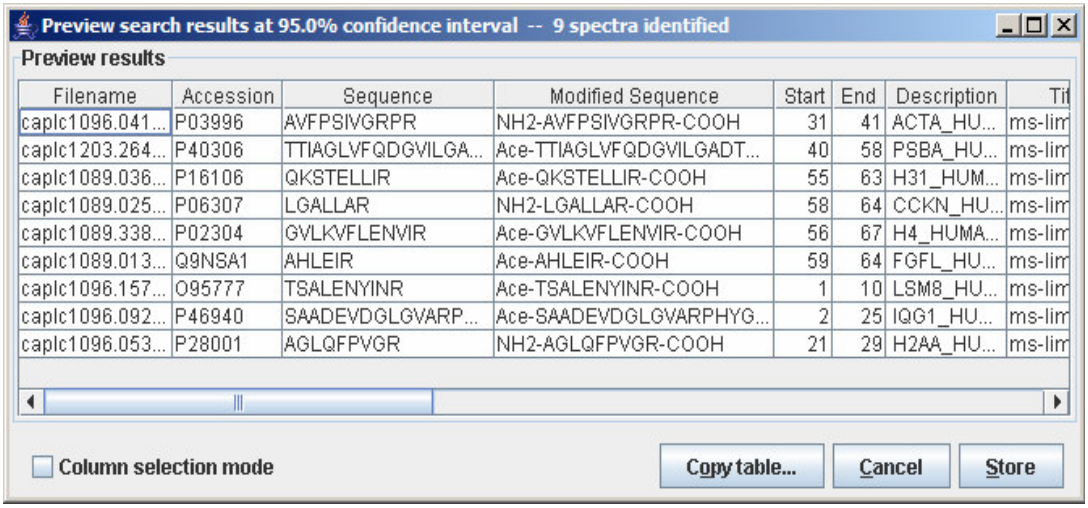


Figure 41: (a) Screenshot of the main screen of IdentificationGUI
(b) Screenshot of the visualization of parsed Mascot results.

The tree view on the left shows the actual queries stored in the Mascot Daemon database, with the results as the leaves of the tree. Selecting queries (or even individual results) in the tree displays the details of the corresponding results on the right-hand table. These details are also directly derived from information stored by Mascot Daemon in its database. In this table, the user can then select the results she/he wants to parse. The identity threshold score applied for the extraction of identifications can be generated according to a user-defined confidence interval. The default is set at the Mascot standard of 95% confidence (allowing for a maximum of 5% false positive identifications). Pressing the 'Preview' button will subsequently cause the application to automatically retrieve all results and parse them according to the specified confidence interval. The user is kept updated on the proceedings by a progress bar. When all results have been parsed, a table is displayed (figure 41(b)) which visualizes the identified peak lists. This table can be manipulated by the user to suit her/his data presentation requirements[53]. Also note that the presented table is at this time not yet stored in the database. The table is meant as a preview, allowing the user decide whether the data should be stored in the ms_lims database at all, or allowing a preview of parts of the data that are already completed while another part is still being processed. It is also possible at this point to export this data to the clipboard for importing into external data analysis software. When the user decides that the presented data can be stored in the database, she/he needs only to click the 'Store' button. The proceedings are once again shown by a progress bar and numerical output of the operations performed is presented to the user in the usual manner (not shown).
An interesting aspect of the data processing performed by IdentificationGUI in addition to the identification performed by Mascot concerns the step from protein identification to peptide identification. IdentificationGUI processes the identified peptides in such a way that, whenever a degenerate peptide is encountered, the list of matching protein identifiers for that peptide is compared with a list of already known protein identifiers (both from the ms_lims database and the processing in progress). As soon as a match is found in this list of 'known' proteins, that protein accession number is selected as 'primary' protein accession number and all other accession numbers are considered 'isoforms'[54]. This implicitly allows for the selection of a minimal explanatory set of protein identifiers as primary accession numbers while maintaining the complete list of matching accession numbers for future reference.

2.2.6. Applications for data analysis

The previous two sections showed initial data gathering and processing in ms_lims. Once the processed data is available in the database however, users will want to work with the

[53] This manipulation can take the form of sorting on a certain column (ascending or descending), resizing columns, moving columns around and clicking certain columns for retrieval of additional information from third-party resources on the web (e.g. clicking the protein accession number retrieves the detailed information from the parent database for that protein).
[54] Note that the usage of the term 'isoform' here is rather specifically aimed at proteins that can all explain a given peptide sequence and that it has no relationship with isoforms as defined by genetic, posttranscriptional or posttranslational events.

results. This analysis stage is the most heterogeneous since many different questions will be asked of the data and it can therefore not be captured by a single application. For ms_lims it was decided to provide specialized applications for specific data querying tasks rather than one big application. Two applications will be discussed in detail here. The first of these is the ProjectAnalyzer, which is a suite of modular tools for generic data retrieval and analysis. The second is a DiffAnalysisGUI, a highly specialized application for the study of peptide-centric proteomics experiments using differentially labeled samples. Differential labeling for these latter experiments is typically achieved through the incorporation of stable isotopes, either during tryptic digestion [Staes 2004, Van Damme 2005] or cell culture [Ong 2002].

2.2.6.1. ProjectAnalyzer application

This application is a control center for a suite of general-purpose analysis modules built around a common framework. A screenshot of the main interface is shown in figure 42.

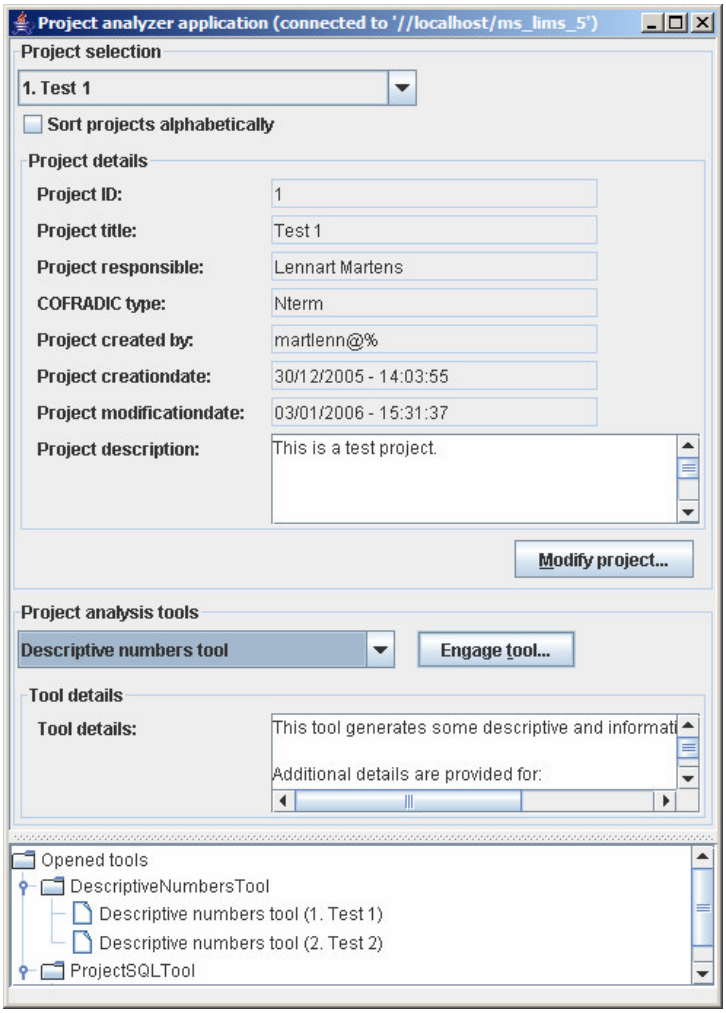


Figure 42: Screenshot of the main ProjectAnalyzer application interface.

The top panel of the interface allows project selection and review or adaptation of the selected projects details. The middle panel allows the selection of an analysis tool from the suite to apply to the selected project and the bottom panel displays all currently opened project analysis tools in a tree view. This bottom component will organize the opened analysis tools by their type, with the leaves corresponding to specific instances of this type that are connected to specific projects. Double-clicking any one of the leaves will also focus the selected analyzer tool on the screen.
These analyzer tools are all built around a framework centered on the *ProjectAnalyzerTool* interface, the UML diagram[55] of which is represented in figure 43.

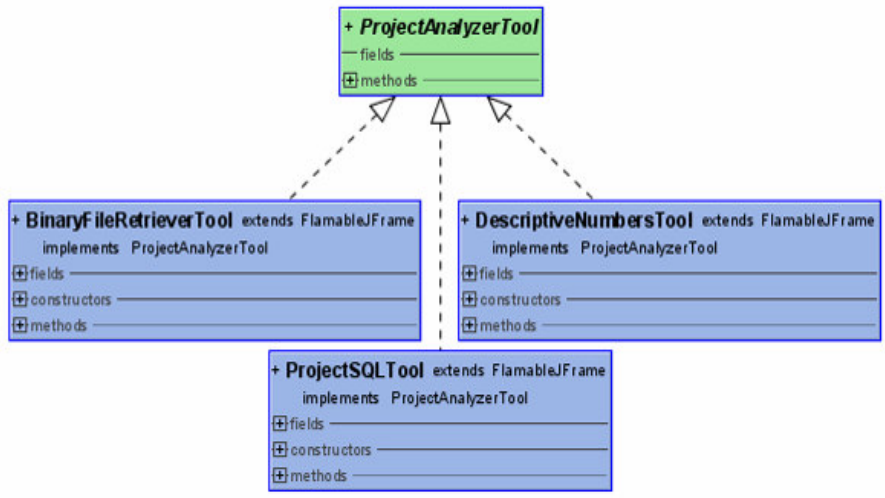


Figure 43: UML diagram of the ProjectAnalyzerTool framework in ms_lims.

When the ProjectAnalyzer application is started, it first queries the user for database connection parameters. After the connection is established, it will proceed to load all projects from the database as well as the contents of the 'projectanalyzertool' table. This table contains a descriptive representation of all known implementations of the *ProjectAnalyzerTool* interface. This dynamic class loading via a common framework is again reminiscent of the system discussed in section 2.2.4 for the retrieval of instrument-specific *SpectrumStorageEngine* implementations. Again, the advantage is clear: it allows third party additions to integrate quickly and seamlessly in a single usage and management interface.
There are three default implementations provided with the latest ms_lims version (ms_lims 5.0): *BinaryFileRetrieverTool*, *DescriptiveNumbersTool* and *ProjectSQLTool*. The *DescriptiveNumbersTool* gives a numerical overview of the project and the *ProjectSQLTool* allows the user to run a set of predefined queries against the selected project. A screenshot of the latter tool is presented in figure 44. A linked-in component called *SpectrumPanel* is available in many ms_lims applications and will enable interactive exploration of a peak lists. Clicking a field in the 'l_spectrumfileid' column in the results table will pop up this component and initialize it with the corresponding spectrum. A screenshot is shown in figure 45.

[55] This UML diagram is not a conceptual model; it is automatically generated on the basis of the actual ms_lims code itself.

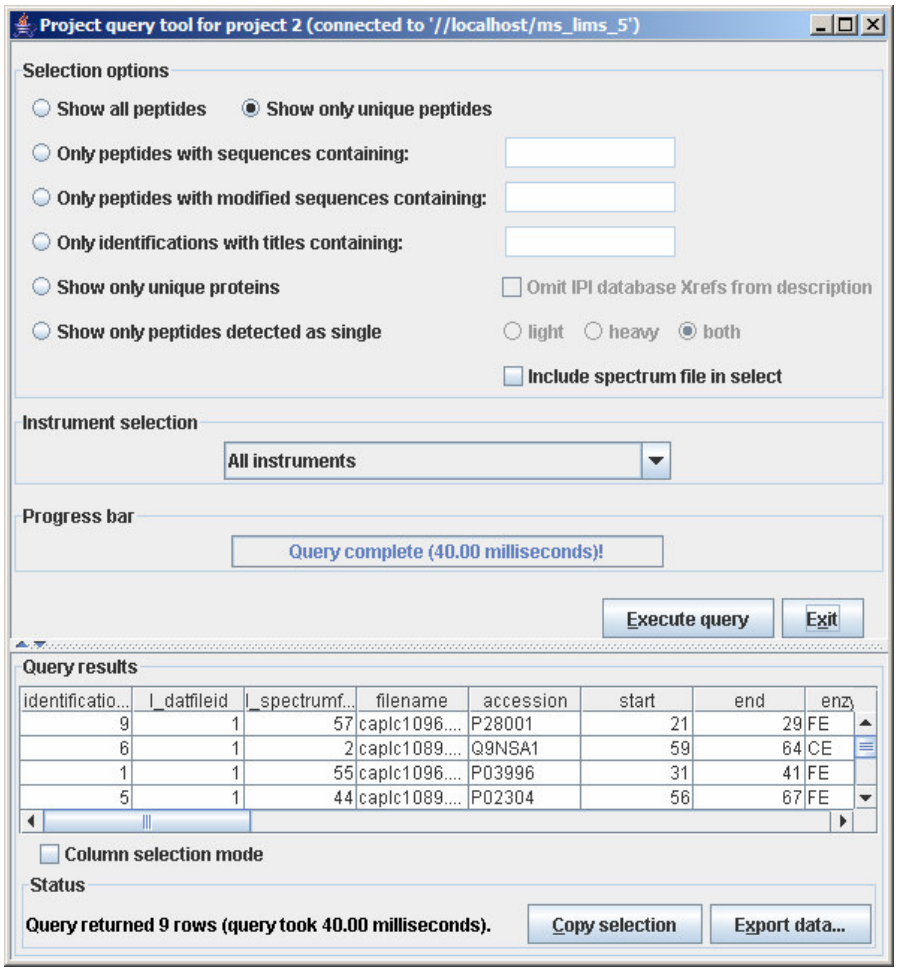


Figure 44: Screenshot of the ProjectSQLTool interface.

Apart form this predefined query selection tool, ms_lims also comes bundled with a generic SQL tool called GenericQuery. This application can connect to any Java DataBase Connectivity (JDBC) compatible database and allows the execution of any kind of query. This way, beginning or non-expert users can quickly become productive using the predefined queries in *ProjectSQLTool* and expert users can maximize their knowledge of the database and of the SQL language by interrogating the database directly. Finally, the generated data access layer delivers a well documented Application Programmers Interface (API) that developers can wield to their advantage when designing their own third-party extensions to the ms_lims system.
The usage of the GNU GPL license (see section 2.2.7 below) guarantees that such third-party efforts will be released back to all potential users, gently enforcing the collaborative aspect of software development for the proteomics community.

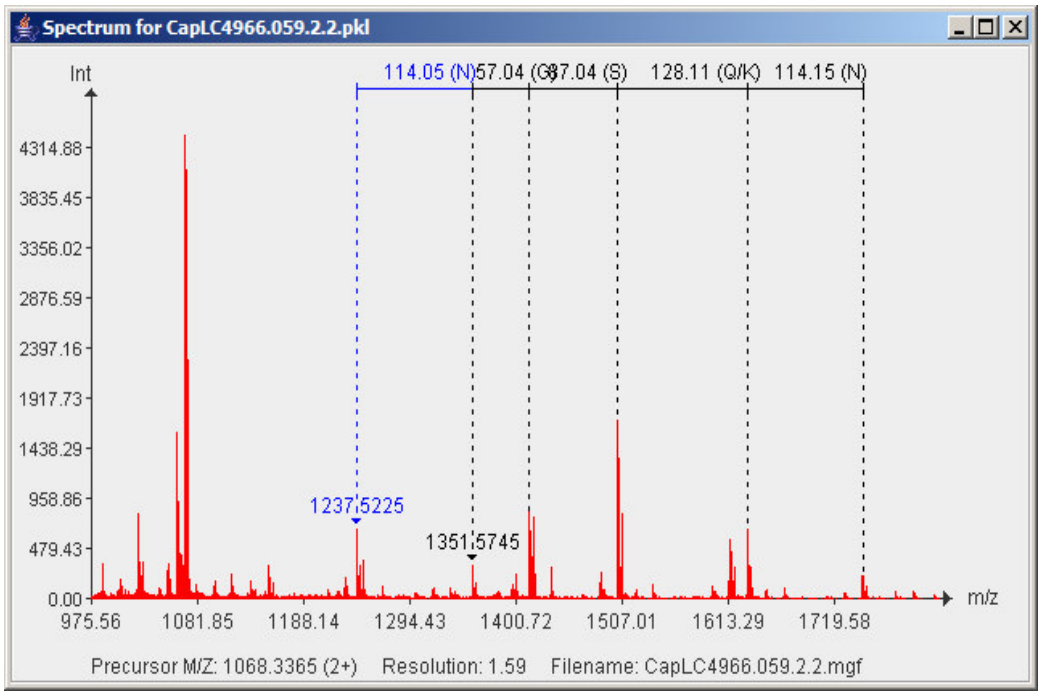


Figure 45: Screenshot of the SpectrumPanel component.

2.2.6.2. DiffAnalysisGUI application

The analysis of differentially labeled proteome samples yields a good example of a complex set of processing steps which need to be performed automatically in order for the scientist to make sense of the mass of data acquired during the experiments. DiffAnalysisGUI presents such an automated processing and analysis application for differential proteomics studies based on isotopically labeled, identified peptides. The 'identification' table in the ms_lims database contains two columns that hold differential data: 'light_isotope' and 'heavy_isotope'. These columns will contain the intensity of the signal for the peptide with the light and heavy isotopic label, respectively. The workflow of DiffAnalysisGUI will be discussed starting from the main interface of the program (figure 46).

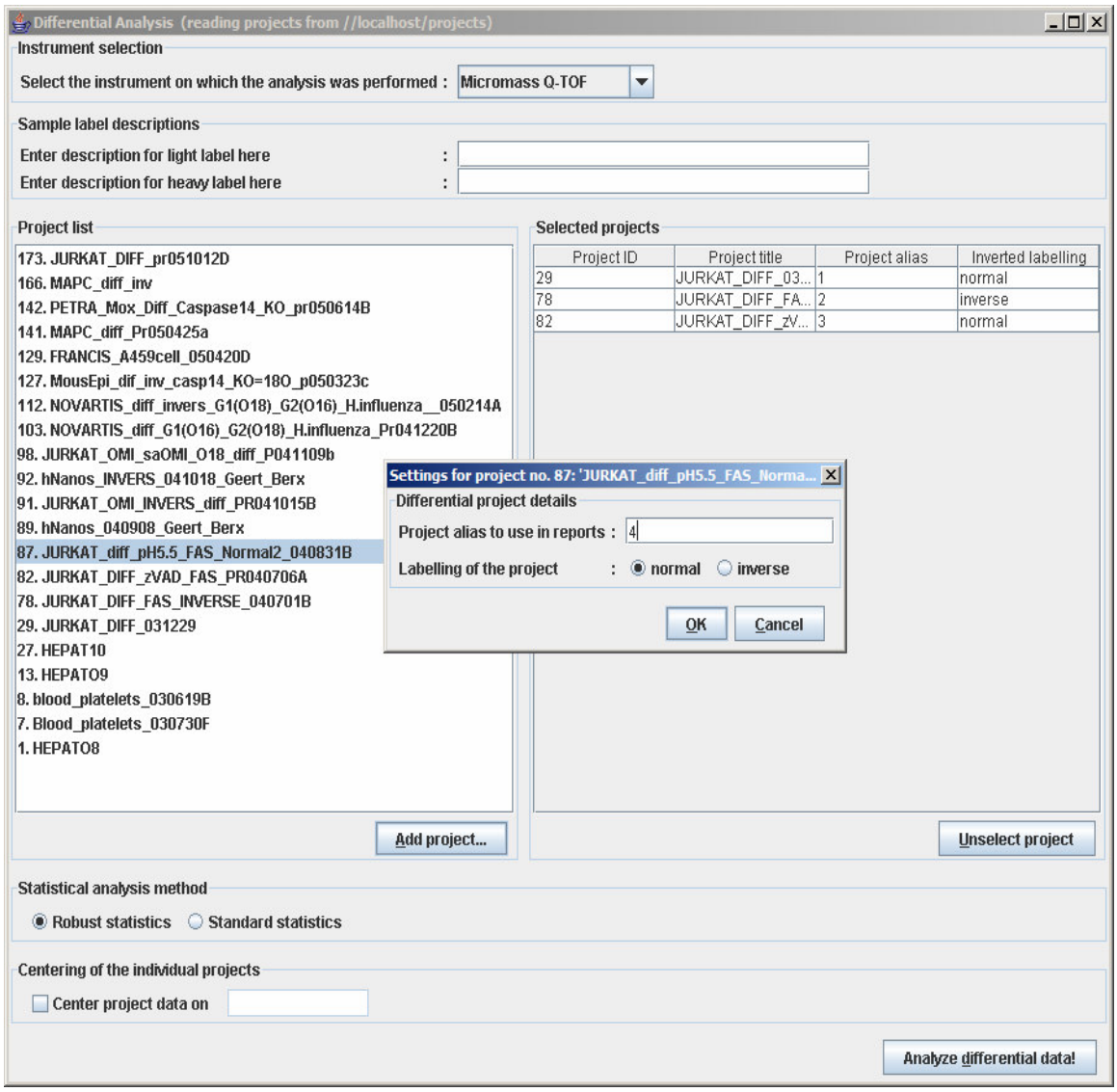


Figure 46: Main interface for the DiffAnalysisGUI application.

Upon starting the program, the user is prompted to provide database connection details. When a connection has been established, the software will proceed to query the database about all known calibrated instruments and all projects that contain differential data. An instrument is considered calibrated when measurements of the two-base logarithms of 1/1 ratios have been performed on this instrument to establish the standard deviation of the detector response and this value has been entered in the 'differential_calibration' column of the 'instrument' table. A project contains differential data as soon as one or more identified peptides for that project have non-null 'light_isotope' and 'heavy_isotope' columns in the 'identification' table.

The top panel of the screen presents a dropdown containing all differentially calibrated instruments. As the standard deviation for each instrument is likely to differ, DiffAnalysisGUI only supports the analysis of differential identifications derived from a single instrument. The next panel allows the user to provide a descriptive label for the

sample that was labeled with the light isotope and one for the heavy isotope[56]. This label will be used throughout to facilitate interpretation of the results. The center panel is split in two sections, the leftmost of which contains a list of all projects containing differential data and the rightmost one showing those projects currently selected for differential analysis. Superimposed on these lists a dialog can be seen in the screenshot that pops up whenever the user adds a project to the currently selected list. This dialog requests a project alias to be entered[57] as well as selection of 'normal' or 'inverse' labeling for this project. Note also that the table on the right of the main interface shows the information entered via this dialog as well as the original project information.

The two bottom panels allow the user to choose the type of statistics that need to be applied and whether re-centering of the data should be performed. The type of statistics is set to 'robust statistics' by default. The influence of the choice of the statistics will be discussed below. The re-centering functionality is particularly important when multiple projects are analyzed together. Since each of the experiments is performed completely separate from all others, there is a very real chance that the labeled/unlabeled mixture is not present in an exact 1/1 ratio in each experiment. Slight deviations to this ratio will yield different distribution centers for each project, in turn artificially broadening the compound distribution. By re-centering each distribution to a specified (expected) ratio, this artificial broadening can be limited while each individual distribution retains its original scale values, thus contributing correctly to the scale of the compound distribution.

The actual analysis flow consists of a few phases, some of which are optional. The user is kept informed of the proceedings through a progress bar.

The first phase concerns the gathering of all data from the database. An SQL query is constructed to report all differential identifications for the selected project(s) and instrument. The second phase is optional. If the user requested a re-centering, the software will calculate the difference between the center of each distribution and the user-specified center and will correct each measured ratio with this difference. The third phase consists of joining all data across all experiments in one compound distribution. This phase also merges spectra that led to the same identified peptide in one 'cluster'. The result is a list of unique (modified) sequences, with each sequence backed by one or more spectra with differential data from the selected instrument. The fourth stage concerns the actual calculation of the statistics. All subsequent calculations are based on the two-base logarithms of the ratios of the clusters, with the ratio for a particular cluster being equal to the average of the ratios of the constituent spectra. According to the user request, the location and scale of the distribution are either calculated by means of the mean and standard deviation ('standard statistics') or through median and Huber scale [Huber 1981] by iterative winsorization at k=1.5[58] ('robust statistics'). The difference between these two approaches revolves around their sensitivity to outliers. The presence of these outliers tends to affect traditional statistics quite heavily yet leaves robust statistics largely unperturbed [Hampel, 1993]. Since differential experiments are typically performed because outliers are expected, robust statistics were chosen to be the default setting. We will designate the location estimator as μ_{hat} and the scale estimator as $s_{[p]}$.

[56] Example descriptive labels could be 'control' and 'diseased'.
[57] Note that plain numbers are used here. More verbose aliasses are of course also possible.
[58] Convergence was set at 10^{-6}.

As the population scale is estimated with high efficiency by the Huber scale, we simply consider:

$$\sigma_{hat} = s_{[p]}$$

Now let the base-two logarithm of each peptide ratio be $r_{[i]}$.
As standard deviation ($s_{[i]}$) for each peptide ratio the calibrated standard deviation for the instrument as represented in the ms_lims database is taken.
Centering is performed by subtracting μ_{hat} from each $r_{[i]}$ to yield a zero-centered measurement ($delta_{[i]}$) and the resulting standard deviation ($s_{[delta,i]}$) was calculated from:

$$delta_{[i]} = r_{[i]} - \mu_{hat}$$

$$s_{[delta,i]} = \sqrt{s_{[i]}^{2} + \sigma_{hat}^{2}}$$

We can now apply a double-sided significance test for each $delta_{[i]}$ at the desired confidence interval to:

$$N \sim (delta_{[i]}, s_{[delta,i]})$$

When the user has selected all relevant projects and verified them and the analysis metadata, she/he can then initiate the actual analysis by pressing the 'Analyze differential data' button. When the analysis completes, the software reports a numerical summary first (figure 47). In the example presented, four projects were subjected to robust statistical analysis and a re-centering was performed on 1.0. Note that the software reports the corrections applied for each separate project. This allows the user to quickly estimate the variability and see potential outliers. In this case, the deviations are very small and of roughly the same magnitude. After having evaluated this summary, clicking 'OK' will bring the user to the result table, shown in figure 48. Note that the title bar of the frame contains the data from the numerical summary. Several entries in the table have been coloured red. This is indicative of outliers within the set of spectra representing this unique sequence. For instance, the highlighted row (identifiable by the light blue background and the red foreground) is composed of combined differential data of eighteen separate spectra all identified to be the same peptide sequence. When clicking this row, the user can verify the outliers on the detail table shown (figure 49).

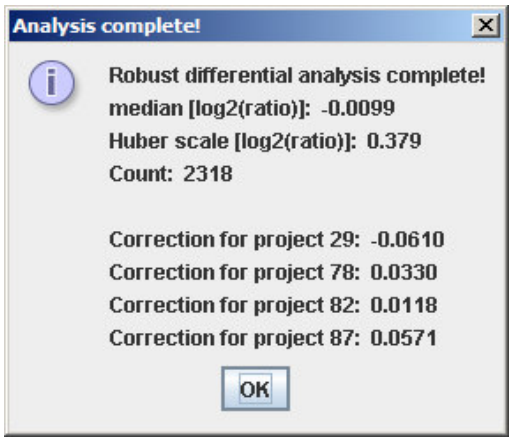
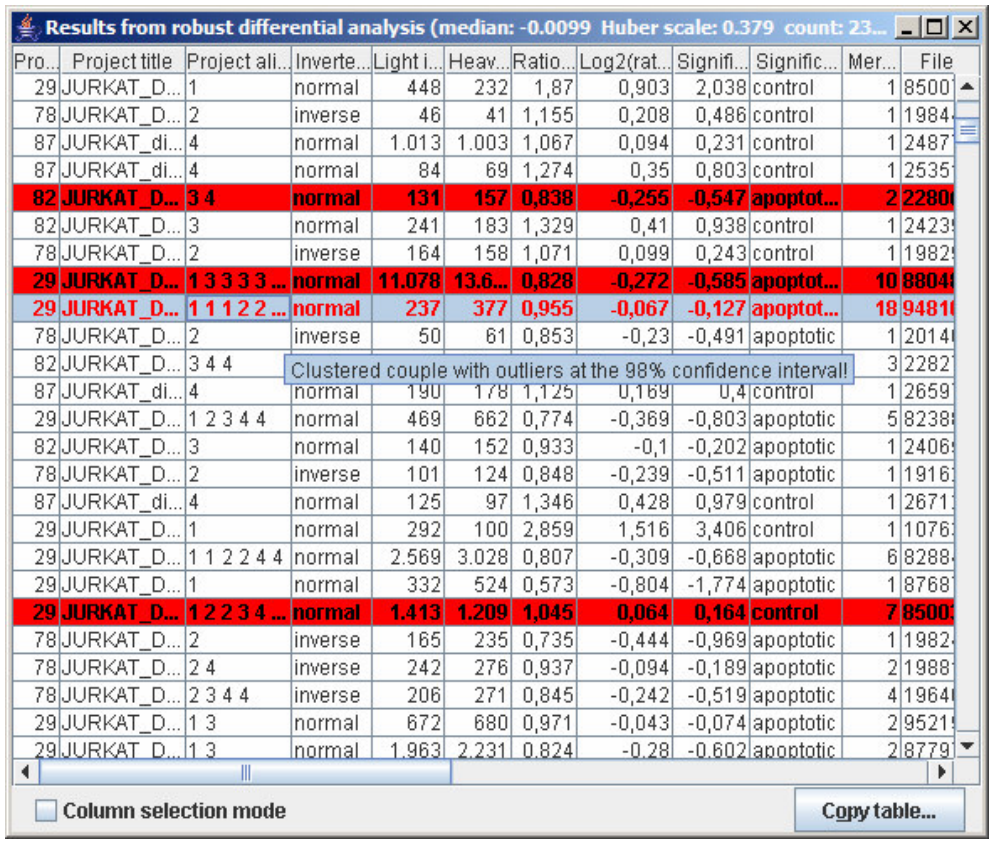


Figure 47: Numerical summary of a four-project robust statistical analysis with a re-centering on 1.0.



Pro...	Project title	Project ali...	Inverte...	Light i...	Heav...	Ratio...	Log2(rat...	Signifi...	Signific...	Mer...	File
29	JURKAT_D...	1	normal	448	232	1,87	0,903	2,038	control	1	8500
78	JURKAT_D...	2	inverse	46	41	1,155	0,208	0,486	control	1	1984
87	JURKAT_di...	4	normal	1.013	1.003	1,067	0,094	0,231	control	1	2487
87	JURKAT_di...	4	normal	84	69	1,274	0,35	0,803	control	1	2535
82	**JURKAT_D...**	**3 4**	**normal**	**131**	**157**	**0,838**	**-0,255**	**-0,547**	**apoptot...**	**2**	**2280**
82	JURKAT_D...	3	normal	241	183	1,329	0,41	0,938	control	1	2423
78	JURKAT_D...	2	inverse	164	158	1,071	0,099	0,243	control	1	1982
29	**JURKAT_D...**	**1 3 3 3 3 ...**	**normal**	**11.078**	**13.6...**	**0,828**	**-0,272**	**-0,585**	**apoptot...**	**10**	**8804**
29	**JURKAT_D...**	**1 1 1 2 2 ...**	**normal**	**237**	**377**	**0,955**	**-0,067**	**-0,127**	**apoptot...**	**18**	**9481**
78	JURKAT_D...	2	inverse	50	61	0,853	-0,23	-0,491	apoptotic	1	2014
82	JURKAT_D...	3 4 4								3	2282
87	JURKAT_di...	4	normal	190	178	1,125	0,169	0,4	control	1	2659
29	JURKAT_D...	1 2 3 4 4	normal	469	662	0,774	-0,369	-0,803	apoptotic	5	8238
82	JURKAT_D...	3	normal	140	152	0,933	-0,1	-0,202	apoptotic	1	2406
78	JURKAT_D...	2	inverse	101	124	0,848	-0,239	-0,511	apoptotic	1	1916
87	JURKAT_di...	4	normal	125	97	1,346	0,428	0,979	control	1	2671
29	JURKAT_D...	1	normal	292	100	2,859	1,516	3,406	control	1	1076
29	JURKAT_D...	1 1 2 2 4 4	normal	2.569	3.028	0,807	-0,309	-0,668	apoptotic	6	8288
29	JURKAT_D...	1	normal	332	524	0,573	-0,804	-1,774	apoptotic	1	8768
29	**JURKAT_D...**	**1 2 2 3 4 ...**	**normal**	**1.413**	**1.209**	**1,045**	**0,064**	**0,164**	**control**	**7**	**8500**
78	JURKAT_D...	2	inverse	165	235	0,735	-0,444	-0,969	apoptotic	1	1982
78	JURKAT_D...	2 4	inverse	242	276	0,937	-0,094	-0,189	apoptotic	2	1988
78	JURKAT_D...	2 3 4 4	inverse	206	271	0,845	-0,242	-0,519	apoptotic	4	1964
29	JURKAT_D...	1 3	normal	672	680	0,971	-0,043	-0,074	apoptotic	2	9521
29	JURKAT_D...	1 3	normal	1.963	2.231	0,824	-0,28	-0,602	apoptotic	2	8779

Figure 48: Result table for the statistical analysis of four re-centered projects.

This detailed view contains the individual spectra that were identified and their details. Colour highlighting is used here as well to show the spectra that are outliers. Yellow is used for outliers at the 95% confidence interval, red for those that still classify as outliers at the 98% interval. Tool tips also show up for these rows with this information.
The most interesting column of this table is the ‘Significance’ column, which expresses the deviation of the measurement from the population center as a product of the

population scale. The user can now apply any desired confidence interval in order to look for outliers (e.g.: finding only those rows with an absolute 'Significance' greater than 1.96 will deliver the outliers at the 95% confidence interval).

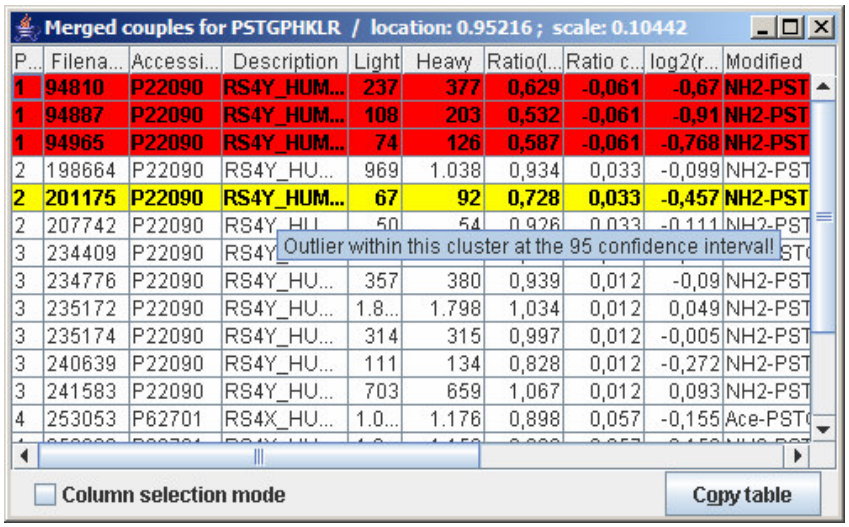


Figure 49: Detailed view on all merged spectra for one sequence.

2.2.7. License and availability

The ms_lims software has been released as open source software under the GNU General Public License (GPL) (http://www.gnu.org/licenses/licenses.html#GPL) and is freely available in both source and binary versions from http://genesis.UGent.be/ms_lims.

2.2.8. Published applications of ms_lims in proteomics experiments

The ms_lims system has been instrumental in data management and analysis alike for nearly all of the papers presented in this dissertation. The papers included below thus only present a sample of the behind-the-scenes impact of the ms_lims software.

2.2.8.1. Cysteine COFRADIC proteome of human blood platelets

Reversible labeling of cysteine-containing peptides allows their specific chromatographic isolation for non-gel proteome studies

Kris Gevaert, Bart Ghesquière, An Staes, Lennart Martens, Jozef Van Damme, Grégoire R. Thomas and Joël Vandekerckhove

Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University and Flanders Interuniversity Institute for Biotechnology, Ghent, Belgium

We report upon a novel procedure to specifically isolate cysteine-containing peptides from a complex peptide mixture. Cysteines are converted to hydrophobic residues by mixed disulfide formation with Ellman's reagent. Proteins are subsequently digested with trypsin and the generated peptide mixture is a first time fractionated by reverse-phase high-performance liquid chromatography. Cysteinyl-peptides are isolated out of each primary fraction by a reduction step followed by a secondary peptide separation on the same column, performed under identical conditions as for the primary separation. The reducing agent removes the covalently attached group from the cysteine side chain, making cysteine-peptides more hydrophilic and, thereby, such peptides can be specifically collected during the secondary separation and are finally used to identify their precursor proteins using automated liquid chromatography tandem mass spectrometry. We show that this procedure efficiently isolates cysteine-peptides, making the sample mixture less complex for further analysis. This method was applied for the analysis of the proteomes of human platelets and enriched human plasma. In both proteomes, a significant number of low abundance proteins were identified next to extremely abundant ones. A dynamic range for protein identification spanning 4–5 orders of magnitude is demonstrated.

Received:	16/7/03
Revised:	18/8/03
Accepted:	1/9/03

Keywords: Cysteine-peptides / Mass spectrometry / Non-gel proteomics

1 Introduction

Several procedures for high-throughput proteome studies have been developed over the past few years. The majority of these commence with a proteolytic digestion of an isolated protein mixture and use peptide-related mass spectrometric (MS) data to identify the precursor proteins and monitor fluxes of their expression levels. The peptide mixture is either analyzed as a whole, or specific sets of peptides are isolated prior to analysis. Because of the enormous complexity of the peptide mixture, the first approach requires a rather thorough separation of this mixture prior to (tandem) MS analysis [1–3]. The second approach is centered around the affinity-based isolation of representative peptides out of a proteome digest and solely uses these to identify the corresponding proteins. Generally, representative peptides are peptides that contain rare amino acids that are well distributed over a given proteome so that the chance that every expressed protein is finally represented by at least one peptide is high. Affinity-based techniques have been described to isolate peptides containing cysteine [4–7], methionine [7, 8], histidine [9], phosphorylated residues [10–12], and *N*-linked carbohydrates [13] for further proteome studies. In all these methods, a general approach for differential, non-gel proteomics involves the incorporation of stable, heavy isotopes in one of the mixtures and MS analysis of the ratio of 'light' and 'heavy' peptides.

Correspondence: Dr. Kris Gevaert, Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University and Flanders Interuniversity Institute for Biotechnology (VIB09), A. Baertsoenkaai 3, B-9000 Ghent, Belgium
E-mail: kris.gevaert@UGent.be
Fax: +32-92649484

Abbreviations: COFRADIC™, combined fractional diagonal chromatography; **DTNB**, 5,5'-dithiobis(2-nitrobenzoic acid); **TCEP**, tris(2-carboxyethyl)phosphine

Supporting information for this article is available on the WWW under www.proteomics-journal.de

Recently, we have developed a chromatographic method for the isolation of sets of representative peptides out of complex mixtures called combined fractional diagonal chromatography or COFRADIC™. Central in this method is a modification reaction which alters the retention behavior on reverse-phase (RP) columns of specific peptides. This reaction is carried out between two identical chromatographic separations. A peptide mixture is first fractionated and in a number of second separations, sets of altered peptides are isolated for further MS/MS analysis. In one application, methionine peptides are isolated following oxidation to their sulfoxide forms in an acidic medium using hydrogen peroxide.

Since methionine-sulfoxide is more hydrophilic than methionine, peptides carrying the oxidized side-chain elute sooner from the column in the secondary run and can thus be isolated from the bulk of peptides that do not contain methionine [14]. By altering the chemical reactions applied in between the primary and secondary separations, we have recently shown that peptides spanning the amino terminal part of proteins can be specifically isolated and used for non-gel proteomics [15]. Besides protein identification, such an analysis allows global monitoring of *in vivo* modifications occurring on the amino termini of proteins (*e.g.*, blocking of the terminal amine and processing of amino terminal amino acids). Using the COFRADIC™ approach, every peptide that contains an amino acid or a chemical group that can be specifically modified such that its chromatographic properties are altered, can be isolated. Here, we present a COFRADIC™ approach by which cysteine-containing peptides are isolated out of a complete proteome digest and are used to identify their parent proteins. In this new method, the proteins present in a mixture are first reduced and the cysteine residues are modified using Ellman's reagent (5,5-dithiobis(2-nitrobenzoic acid) or DTNB) [16] to form a mixed disulfide product (TNB-cysteine). The hydrophobic TNB group remains covalently bound during protein digestion and RP-HPLC separation and upon its removal using a reducing agent, cysteine-containing peptides become more hydrophilic and can be collected during secondary, identical RP-HPLC runs. Finally, the collected peptides are used to identify their precursor proteins using automated LC-MS/MS analysis. In order to illustrate the potential of this technique for gel-free proteome analysis, we analyzed the proteomes of a human blood platelet preparation and of a human plasma sample from which serum albumin and immunoglobulins were largely depleted. From the generated data it is clear that a significant enrichment of cysteinyl peptides is achieved using the developed sorting procedure. The obtained reduction of the sample's complexity furthermore makes it possible to identify proteins of which the concentration spans an interval of 4–5 orders of magnitude, as is demonstrated by the proteome analysis of the plasma sample.

2 Materials end methods

2.1 Isolation of the human platelet proteome

A fresh platelet-rich suspension containing approximately 50×10^9 cells was obtained from the Red Cross Blood Transfusion Centre Oost-Vlaanderen, Ghent, Belgium. This suspension was centrifuged for 10 min at $1000 \times g$ and the pellet was resuspended in 40 mL of washing buffer containing 0.2 g/L KCl, 0.2 g/L $MgCl_2$, 8 g/L NaCl, 1 g/L D-glucose, and 70 mg/L EGTA in 25 mM sodium phosphate buffered at a pH of 7.5. This washing procedure was repeated twice and the final platelets pellet was resuspended in a total volume of 10 mL of wash buffer and lysed by addition of 10 mL of 0.5% Triton X-100 in 25 mM of sodium phosphate buffer (pH 7.5) containing protease inhibitors (Roche Diagnostics, Mannheim, Germany). This suspension was kept for 30 min on ice followed by centrifugation for 10 min at $10\,000 \times g$ in order to remove cellular debris. 500 μL of the obtained protein mixture (corresponding to 1.25×10^9 platelets) was desalted on a NAP™-5 column (Sephadex G25 DNA Grade from Amersham Biosciences, Uppsala, Sweden) in 2 M guanidinium hydrochloride in 50 mM Tris·HCl (pH 8.7). The desalted proteins were collected in 1 mL and stored at −80°C until further use.

2.2 Isolation of the human plasma proteome

Human plasma was also obtained from the Red Cross Blood Transfusion Centre Oost-Vlaanderen. It was first centrifuged for 10 min at $1000 \times g$ to remove cells that were present in the sample. From 100 μL of cell-free plasma serum albumin was removed using the Montage Albumin Depletion Kit from Millipore (Billerica, MA, USA) according to the manufacturer's instructions. The albumin-depleted plasma was then further incubated with 30 μL of Protein A Sepharose™ 4 Fast Flow beads (Amersham Biosciences) to remove immunoglobulins. The initial plasma sample was finally collected in 600 μL. Of this, 500 μL was desalted on a NAP™-5 column as described above for the platelet proteins and collected in a 1 mL volume.

2.3 Modification of cysteine residues

The desalted protein mixtures were lyophilized and redissolved in 500 μL of 50 mM tris(2-carboxyethyl)phosphine (TCEP) (Pierce, Rockford, Il, USA). Proteins were reduced

in this solution for 1 h at 37°C and were then desalted as described above on a NAP™-5 column. This protein mixture was dried and the cysteine residues were modified by adding 500 μL of 10 mM Ellman's reagent (Fluka, Steinheim, Germany) in 100 mM Tris·HCl pH 8.7 for 1 h at 37°C. The modified protein mixture was finally desalted on a NAP™-5 column in 1 mL of 10 mM Tris·HCl at pH 8.7 and concentrated to half its volume by vacuum drying. Prior to digestion, the protein mixture was boiled for 10 min and cooled on ice for 15 min. Five microgram of sequencing-grade modified trypsin dissolved in 25 μL of 1 mM of acetic acid (Promega, Madison, WI, USA) was added and the digestion proceeded overnight at 37°C and was stopped by acidification with 50 μL of 10% trifluoroacetic acid (TFA). The generated peptide mixture was stored at −20°C prior to RP-HPLC separation.

2.4 COFRADIC™-based isolation of cysteine-containing peptides

For the analysis of the platelet proteome, 100 μL of the tryptic digest was used; this corresponds to the protein material present in about 2×10^8 platelets. Likewise, 400 μL of the plasma proteome digest was used for proteome analysis, which corresponds to about 67 μL of the original plasma sample. Prior to RP-HPLC analysis, the proteome digests were centrifuged to remove any insoluble material and all methionine residues were converted to their sulfoxide counterparts by adding fresh hydrogen peroxide to a final concentration of 0.5% w/v to the peptide mixtures and incubating it for 30 min at 30°C [14]. This step is necessary since during the isolation procedure, methionine residues may become oxidized and peptides containing such residues will show a hydrophilic shift and will thus also be isolated. Immediately following the oxidation step, the peptide mixtures were injected onto a narrow-bore reverse-phase ZORBAX 300SB-C18 column (2.1 mm id × 150 mm; Agilent Technologies, Waldbronn, Germany) coupled to an Agilent 1100 Series Capillary LC-system controlled by the Agilent ChemStation software modules. Following sample injection, a binary solvent gradient was applied at a controlled constant flow rate of 80 μL/min (microflow mode). The column was first washed with 0.1% TFA in water (Baker HPLC analyzed; Mallinckrodt Baker B.V., Deventer, The Netherlands) (solvent A) for 10 min, followed by a linear gradient to 70% acetonitrile (Baker HPLC analyzed) in 0.1% TFA (solvent B) over 100 min. This RP-HPLC separation is referred to as the primary run. Eluting peptides were collected from 40 min (30% of solvent B) onwards in a total of 48 fractions of 1 min (or 80 μL) in a microtiterplate using the Agilent 1100 Series fraction collector (see Table 1). Primary fractions that were separated by 16 min were pooled (Table 1) and dried in a centrifugal vacuum concentrator. These dried fractions were redissolved in 70 μL of 10 mM Tris·HCl (pH 8.7) and 30 μL freshly prepared 50 mM TCEP was added to remove the 3-carboxy-4-nitrophenylthiol moiety from the cysteines. This reduction reaction proceeded for 1 h at 37°C and was stopped by acidification with TFA. The vials containing the reduced and pooled primary fractions were then placed in the Agilent 1100 Series Well-plate sampler. Per secondary run, 95 μL of each treated primary fraction was injected onto the RP-HPLC column and the peptides were separated using the same buffer gradient as applied during the primary run. In these conditions, cysteine-containing peptides elute in a time window of 10 min to 3 min in front of the unaltered (cysteine-free) peptides and were collected in six subfractions per primary fraction (Table 1). Subfractions with an identical subscript and derived from the same secondary run were pooled, dried and put at −20°C until LC-MS/MS analysis was started.

Table 1. Scheme indicating the collection times for the cysteine-containing peptides during COFRADIC™ analysis

	Primary fractions	Elution of primary fractions (min)	Elution of secondary fractions (min)
A	16, 32, 48	55–56, 71–72, 87–88	45–52, 61–68, 77–84
B	15, 31, 47	54–55, 70–71, 86–87	44–51, 60–67, 76–83
C	14, 30, 46	53–54, 69–70, 85–86	43–50, 59–66, 75–82
D	13, 29, 45	52–53, 68–69, 84–85	42–49, 58–65, 74–81
E	12, 28, 44	51–52, 67–68, 83–84	41–48, 57–64, 73–80
F	11, 27, 43	50–51, 66–67, 82–83	40–47, 56–63, 72–79
G	10, 26, 42	49–50, 65–66, 81–82	39–46, 55–62, 71–78
H	9, 25, 41	48–49, 64–65, 80–81	38–45, 54–61, 70–77
I	8, 24, 40	47–48, 63–64, 79–80	37–44, 53–60, 69–76
J	7, 23, 39	46–47, 62–63, 78–79	36–43, 52–59, 68–75
K	6, 22, 38	45–46, 61–62, 77–78	35–42, 51–58, 67–74
L	5, 21, 37	44–45, 60–61, 76–77	34–41, 50–57, 66–73
M	4, 20, 36	43–44, 59–60, 75–76	33–40, 49–56, 65–72
N	3, 19, 35	42–43, 58–59, 74–75	32–39, 48–55, 64–71
O	2, 18, 34	41–42, 57–58, 73–74	31–38, 47–54, 63–70
P	1, 17, 33	40–41, 56–57, 72–73	30–37, 46–53, 62–69

During the primary run, 48 distinct fractions are obtained. Per secondary run (A–P), 3 primary fractions (indicated in the second column) which are separated by 16 min (elution times are given in the third column) are pooled and reduced using TCEP. The cysteine-containing peptides shift out of these primary collection intervals and elute in a time frame that starts 10 min before the elution of the primary fraction and which lasts for 7 min (see fourth column). Typically, the secondary fractions are collected in 6 subfractions and per secondary run, subfractions that are identically indexed are pooled, dried, and applied for final LC-MS/MS analysis.

2.5 LC-MS/MS analysis and peptide identification

Automated LC-MS/MS identification of cysteine-containing peptides was done essentially as described previously [14]. Pooled and dried peptides were redissolved in 20 μL of solvent A' consisting of 0.1% formic acid and 2% acetonitrile in water. 10 μL was loaded on a 0.3 mm ID × 5 mm trapping column (PepMap, LC Packings, Amsterdam, The Netherlands) at a flow rate of 20 μL/min (total loading time of 5 min) using a CapLC System (Micromass UK Limited, Cheshire, UK). By switching the stream valve, the trapping column is back-flushed with a binary solvent gradient, which is started simultaneously with the injection cycle, and the sample is thereby loaded on a nano-scale reverse-phase C18 column (0.75 μm id × 150 mm PepMap™ column; LC Packings). The solvent delivery system was set at a constant flow of 5 μL/min and using a 1/25 flow splitter, 200 nL/min of solvent was directed through the nano-column. Peptides were eluted from the stationary phase with a gradient from 0% to 100% solvent B' (70% acetonitrile in 0.1% formic acid) created over a 25 min period. The outlet of the nano-column was in-line connected to a distal metal-coated fused-silica PicoTip™ needle (PicoTip™ FS360-20-10-D-C7; New Objective, Woburn, MA, USA), placed in front of the inlet of a Q-TOF1 mass spectrometer (Micromass UK Limited). Automated data-dependent acquisition with the Q-TOF mass spectrometer was initiated 15 min after the stream valve was switched. The acquisition parameters were chosen such that only doubly and triply charged ions were selected for fragmentation and for each fragmented peptide spectra were obtained over a period of 8 s. The stream valve was switched back 51 min after the start of the injection cycle.

The acquired collision-induced dissociation (CID)-spectra were automatically converted to a MASCOT [17] acceptable format using Proteinlynx available in the Micromass MassLynx software (Version 3.4). Three different sequence databases were used in sequence to identify the isolated peptides. MASCOT searched first in the SWISS-PROT database, restricting taxonomy to human proteins (ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb/fasta/sprot.fas.gz), then in NCBI's nonredundant protein database (ftp://ftp.ncbi.nih.gov/blast/db/nr.tar.gz) and finally in an in-house created database holding all possible cysteine-containing peptide sequences derived from human proteins present in the SWISS-PROT database. The only restriction applied to these peptides was that their mass had to fall within the range of 600 to 4000 Da. To avoid the possibility that one MS/MS-spectrum was linked to identical sequences stored in the three different databases, the follow-up search option that is available in the MASCOT daemon tool was used in such a way that only the MS/MS-spectra that were not identified in one database, were used to search in the larger databases. The following MASCOT search parameters were set: enzyme: trypsin, maximum number of missed cleavages: 1, variable modification: deamidation (N and Q), oxidation (M), acetyl (*N*-terminus of protein) and pyroglutamate formation (*N*-terminal Q), peptide tolerance: 0.3 Da, MS/MS tolerance: 0.3 Da and peptide charge: 2+/3+. Searches in the third, peptide-centric database were not restricted by an enzyme setting. The DAT result files of MASCOT were automatically queried using in-house developed software tools and only MS/MS-spectra that were identified by a score that exceeded the identity threshold score of MASCOT were retained. The identified peptides were automatically stored in a MySQL relational database in which links were made to their MS/MS-spectra and precursor proteins.

3 Results

3.1 COFRADIC™-based isolation of cysteine-containing peptides

Central in COFRADIC™ is the use of a chemical modification step that alters the chromatographic behavior of a subset of representative peptides between two identical RP-HPLC separations such that targeted peptides can be isolated [14, 15]. For the COFRADIC™-based isolation of cysteinyl-peptides, cysteine residues were modified prior to protein digestion. This is routinely done in MS-based protein analyses since it denatures proteins and makes more sites accessible for proteolytic cleavage, and thus leads to a higher protein sequence coverage [18]. In the work presented here, following a reduction step, cysteine residues are modified by DTNB (Ellman's reagent) [16]. This reaction renders cysteines highly hydrophobic, since a mixed disulfide between the side chain of cysteine and a thionitrobenzoic acid (TNB) group is generated (Fig. 1).

We use Ellman's reagent for two important reasons: (i) this reaction is known to be highly quantitative (*e.g.*, [19], [20]), and (ii) the TNB-group can be easily removed prior to the secondary separation using a reducing agent. Following this modification reaction, proteins are digested and fractionated a first time by RP-HPLC. In what follows, we refer to this separation as the primary run and typically, 48 primary fractions are collected (Fig. 2A). These fractions contain two types of peptides; those containing at least one TNB-altered cysteine and those that are free of cysteine residues. Upon treatment of a primary fraction with a reducing agent – here we use the strong reducing

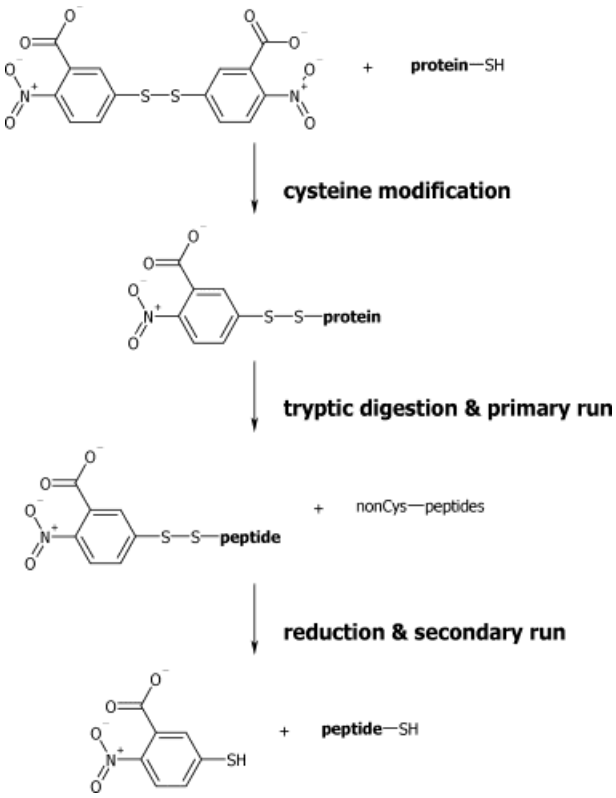


Figure 1. Chemical reactions employed for the COFRADIC™-based isolation of cysteine-peptides. All the cysteine residues of proteins are modified with Ellman's reagent prior to digestion. Upon tryptic digestion, two sorts of peptides are generated; peptides carrying (an) altered cysteine residue(s) and cysteine-free peptides (nonCys-peptides). This peptide mixture is fractionated by RP-HPLC (primary run). The covalently attached group is removed from the cysteine residues by a simple reduction step, which makes the cysteine-peptides more hydrophilic. Finally, upon a secondary separation, these peptides will shift out of the primary collection interval and can thus be isolated for further analysis.

agent and water-soluble TCEP [21] – the TNB-group is removed from all peptides carrying cysteine(s). When such a reduced fraction is re-run on the same column and under identical chromatographic conditions (the secondary run(s)), a hydrophilic shift of the reduced, cysteine-containing peptides is evoked, whereby these move out of the primary collection interval and can be specifically collected (see Fig. 2B).

We have tested this reversible cysteinyl modification on a number of synthetic peptides and under the conditions used (see Section 2), cysteine-containing peptides typically elute in a time frame between 3 and 10 min prior to their primary collection interval. This indicates that multiple primary fractions can be pooled, reduced and fractionated per secondary run. Hence, the number of secondary runs and thus the total analysis time can be significantly reduced. Here, we combine three primary fractions that are separated by 16 min during the primary run (Table 1) and collect the reduced cysteine-containing peptides of three primary fractions during one secondary run (Fig. 2B). We have noticed that when more than three primary fractions are pooled, peptides shifting from one primary fraction can elute in the time interval during which nonshifting (cysteine-free) peptides from a previous primary fraction elute. For this reason, we pooled three primary fractions per secondary run. The cysteine-containing peptides that are shifted out of one primary fraction are typically collected in six subfractions (see Fig. 2B) and prior to LC-MS/MS analysis, secondary subfractions bearing the same index are pooled and dried; *e.g.*, for the secondary RP-HPLC separation depicted in Fig. 2B, secondary subfractions 9_1, 25_1 and 41_1 are pooled. This implies that 96 LC-MS/MS runs will be performed per proteome analysis using COFRADIC™-isolated cysteinyl-peptides.

3.2 Proteome analysis of human platelets using isolated cysteine-containing peptides

The developed cysteine-COFRADIC™ protocol was applied to identify the proteins present in a proteome preparation of human platelets. The isolated cysteine-containing peptides were analyzed by automated LC-MS/MS and 2665 CID-spectra were obtained. Of these, 611 spectra (or 22.9%) were unambiguously linked to human peptide sequences using the MASCOT database search algorithm. Following database searching, we noticed that 99.2% of the identified spectra were linked to fully tryptic peptides and the remainder (0.8%) were from partially tryptic peptides. 522 of the identified peptides contained at least one cysteine residue, indicating that our procedure has an isolation efficiency of 85.4%. Some peptides were analyzed and identified several times, reducing the number of unique peptide sequences to 362. The list of identified peptides and their corresponding proteins is available as supplemental data (Supplementary Table 1). Note that 63 peptides cannot be linked to a single protein entry since at least one other protein (isoform) carrying the same peptide (see supplementary Table 1) is identified. The obtained data allowed us to identify 163 different proteins.

Proteins that are frequently underrepresented in 2-D gels include proteins with low copy numbers and hydrophobic proteins (*e.g.*, [22]). To obtain a global view of our results, the list of identified proteins was mapped onto the human gene ontology annotation (GOA) database [23]. Patterns were detected using a software tool that allows

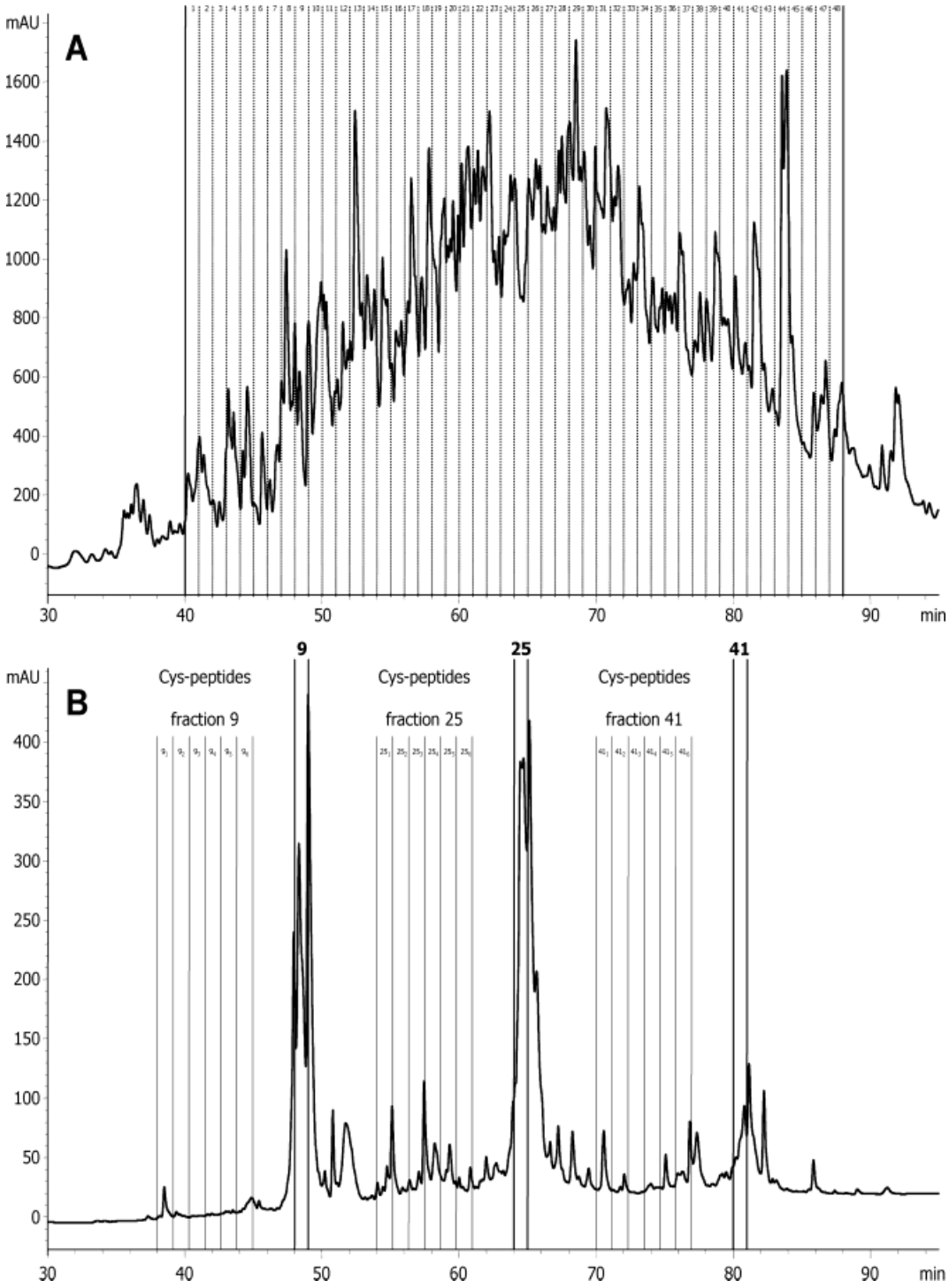


Figure 2. Principle of COFRADIC™ for the isolation of cysteine-peptides. (A) A complex peptide mixture – here, a tryptic digest of human plasma depleted of serum albumin and immunoglobulins –, is first separated by RP-HPLC (the UV-absorbance profile at 214 nm is shown). During the primary run, 48 different primary fractions are obtained. (B) Prior to the secondary run, three fractions that are separated by 16 min (here, fractions numbered 9, 25, and 41) are pooled and treated with a reducing agent, which removes the TNB group from the cysteine residues. When re-run on the same column and under identical chromatographic conditions, cysteine-containing peptides have become more hydrophilic and thus shift out of the primary collection interval and can be collected in an time interval of 10 to 3 min prior to their original elution interval.

visual analysis of the complex data structures used in the GOA database. Based on the biological function and localization of the identified proteins, 3 known protein kinases (the tyrosine kinases Src and Yes and the integrin-linked protein kinase) and 11 proteins that have at least one helix passing through a biological membrane were identified (Table 2). Since estimating the abundances of proteins in gel-free proteome analysis without adding internal standards [24] is difficult, we examined the list of identified proteins in 2-D gel based studies to see whether these selected proteins were previously identified. Only two membrane-spanning proteins have been previously identified in the platelet proteome using 2-D gels [25] (Table 2). Furthermore, none of the kinases or the membrane spanning proteins are stored in the Swiss-2DPAGE (http://ca.expasy.org/ch2d/). Taking together these observations indicate that these proteins are not routinely identified on 2-D gels and may thus be low-abundant (*e.g.*, the protein kinases) or too hydrophobic (*e.g.*, the membrane passing proteins).

Table 2. List of human platelet proteins having at least one membrane spanning-domain

Acc. No.	Protein description
P27105	Erythrocyte band 7 integral membrane protein
P23229	Integrin alpha-6 precursor
P08514	Integrin alpha-IIb precursor (*)
P05556	Integrin beta-1 precursor
P05106	Integrin beta-3 precursor
O00264	Membrane-associated progesterone receptor component 1 (*)
P16284	Platelet endothelial cell adhesion molecule precursor
P07359	Platelet glycoprotein Ib alpha chain precursor
P16671	Platelet glycoprotem IV
P16109	P-selectin precursor
P16615	Sarcoplasmic/endoplasmic reticulum calcium ATPase 2

Following searches in the gene ontology database, 11 different proteins were identified in our platelet proteome that had at least one helix traversing through a biological membrane. Their SWISS-PROT accession numbers and descriptions are given in the first two columns. Proteins that were identified in a previous, 2-D gel based study [25] are indicated with an asterisk.

3.3 Proteome analysis of human plasma depleted from serum albumin and immunoglobulins using isolated cysteine-containing peptides

For the proteome analysis of a human plasma sample we decided to deplete serum albumin and antibodies from the protein mixture as we expected to isolate too many cysteine-containing peptides from these proteins, which would mask peptides originating from minor proteins. These depletion steps were visualized following SDS-PAGE and, as specified by the manufacturer, most but not all of the serum albumin and immunoglobulins were removed; *e.g.*, about one-third of the albumin could not be extracted (data not shown). A total of 6253 MS/MS-spectra was obtained using the cysteine-peptides isolated out of a tryptic digest of 67 μL of the plasma sample. Using MASCOT and restricting the search to human proteins, 1523 spectra were identified (24.4%) of which 1094 (71.8%) contained at least one cysteine residue. In this case, 96.1% of the identified peptides were correctly processed, 3.7% were partially tryptic and only 0.2% (3 peptides in total) were nontryptic. A total of 384 unique peptides sequences were linked to the identified spectra, thereby identifying 102 different proteins (see Table 3). As expected some of the peptides could not be linked to an unique protein sequence and thus identified multiple protein isoforms. This was especially prominent when tryptic peptides from antibodies were isolated and identified. For example, the isolated peptide Ac-SDDTAVYYCAR-COOH could be linked to as many as 428 different protein entries, all being the heavy chain variable region of antibodies!

Table 3. List of proteins identified in a human plasma sample from which albumin and immunoglobulins were depleted

	Description	Acc. No.	LC^2-2D-PAGE	LC^2-MS/MS	Concentration		# of cysteines	# of non-Cys-peptides
1	Fibrinogen alpha/alpha-E chain precursor	P02671			2–4 g/L	[28]	12	6
2	Fibrinogen beta chain precursor	P02675			2–4 g/L	[28]	11	5
3	Fibrinogen betaB 1–118	223130			2–4 g/L	[28]	3	1
4	Fibrinogen gamma chain precursor	P02679			2–4 g/L	[28]	10	2
5	Gamma-fibrinogen chain fragment	577055			2–4 g/L	[28]	0	1
6	Fibrinogen alphaA	223918			2–4 g/L	[28]	6	1

Table 3. Continued

	Description	Acc. No.	LC^2-2D-PAGE	LC^2-MS/MS	Concentration		# of cysteines	# of non-Cys-peptides
7	Serotransferrin precursor	P02787	✓	✓	1.8–2.7 g/L	[28]	38	9
8	Haptoglobin-1 precursor	P00737	✓	✓	0.27–2.47 g/L	[29]	9	2
9	Haptoglobin-2 precursor	P00738	✓	✓	0.27–2.47 g/L	[29]	12	1
10	Complement C3 precursor	P01024	✓	✓	1–2 g/L	[28]	27	5
11	*Apolipoprotein A-I precursor*	*P02647*	✓	✓	*1.27–1.47 g/L*	*[28]*	*0*	*8*
12	*Alpha-1-antitrypsin precursor*	*P01009*	✓	✓	*1.3 g/l*	[28]	*1*	*14*
13	Alpha-1-acid glycoprotein 1 precursor	P02763		✓	0.45–1.28 g/L	[29]	4	1
14	*Alpha-2-macroglobulin precursor*	*P01023*	✓	✓	*1.2 g/L*	*[28]*	*25*	*15*
15	Hemopexin precursor	P02790	✓	✓	0.50–1 g/L	[28]	12	1
16	*Apolipoprotein A-II*	*P02652*	✓	✓	*650–740 mg/L*	*[28]*	*1*	*1*
17	Alpha-1-acid glycoprotein 2 precursor	P19652		✓	~700 mg/L	[30]	5	1
18	*Alpha-2-HS-glycoprotein precursor*	*P02765*	✓	✓	*630 mg/L*	*[28]*	*11*	*2*
19	Ceruloplasmin precursor	P00450	✓		200–600 mg/L	[29]	14	1
20	Alpha-1-antichymotrypsin precursor	P01011	✓	✓	300–600 mg/L	[29]	1	2
21	Complement C4 precursor	P01028	✓	✓	200–600 mg/L	[28]	28	0
22	Complement factor H precursor	P08603	✓	✓	200–600 mg/L	[28]	80	0
23	C4 complement C4d region	13936421			200–600 mg/L	[28]	1	1
24	Vitamin D-binding protein precursor	P02774	✓	✓	400 mg/L	[28]	28	0
25	Vitronectin precursor	P04004	✓	✓	350 mg/L	[28]	14	0
26	Fibronectin	11119231			300 mg/L	[28]	16	0
27	*Transthyretin precursor*	*P02766*	✓		*250 mg/L*	*[28]*	*1*	*2*
28	Complement factor B precursor	P00751	✓	✓	210 mg/L	[28]	23	0
29	Coagulation factor XIII B chain precursor	P05160	✓	✓	210 mg/L	[31]	40	0
30	Plasminogen precursor	P00747	✓	✓	70–200 mg/L	[28]	48	0
31	Beta-2-glycoprotein I precursor	P02749	✓		150–170 mg/L	[32]	22	0
32	Chain L, Antithrombin	1000039	✓		115–160 mg/L	[28]	6	0
33	Antithrombin-III precursor	P01008	✓		115–160 mg/L	[28]	6	0
34	C4b-binding protein alpha chain precursor	P04003	✓	✓	150 mg/L	[28]	36	0
35	Prothrombin precursor	P00734	✓	✓	110 mg/L	[28]	24	0
36	Clusterin precursor	P10909	✓		35–105 mg/L	[28]	10	0
37	AMBP protein precursor	P02760	✓	✓	20–100 mg/L	[28]	16	0
38	Apolipoprotein D	P05090	✓	✓	~100 mg/L	[33]	5	0
39	Kininogen precursor	P01042	✓	✓	~75 mg/L	[28]	18	0
40	Zinc-alpha-2-glycoprotein precursor	P25311	✓		40–75 mg/L	[28]	4	0
41	Angiotensinogen precursor	P01019	✓		~70 mg/L	[34]	4	0
42	Alpha-2-antiplasmin precursor	P08697	✓	✓	69 mg/L	[28]	4	0
43	Heparin cofactor II precursor	P05546	✓	✓	~60 mg/L	[28]	3	1
44	Plasma kallikrein precursor	P03952	✓	✓	35–50 mg/L	[28]	37	0
45	Plasma retinol-binding protein precursor	P02753	✓		46 mg/L	[28]	6	0
46	Complement C1r component precursor	P00736	✓	✓	34 mg/L	[28]	27	0
47	Apolipoprotein C-IV precursor	P55056			1–19 mg/L	[35]	3	1

Table 3. Continued

	Description	Acc. No.	LC^2-2D-PAGE	LC^2-MS/MS	Concentration		# of cysteines	# of non-Cys-peptides
48	Coagulation factor V precursor	P12259			4–10 mg/L	[28]	19	0
49	Sex hormone-binding globulin precursor	P04278	✓		~4 mg/L	[36]	4	0
50	Calgranulin A	P05109	✓		10–100 μg/L	[37]	1	1
51	Complement receptor type 1 precursor	P17927			13–81 μg/L	[28]	120	0
52	55 kDa erythrocyte membrane protein	Q00013			N.F.		4	0
53	N33 protein	Q13454			N.F.		6	1
54	Alpha-1B-glycoprotein precursor	P04217	✓	✓	N.F.		10	0
55	Vascular endothelial growth factor D precursor	O43915			N.F.		9	1
56	Similar to katanin p60	12653659			N.F.		6	0
57	Complement-activating component of Ra-reactive factor precursor	P48740	✓		N.F.		28	0
58	Afamin precursor	P43652	✓	✓	N.F.		34	0
59	Peptidoglycan recognition protein L precursor	15705411			N.F.		10	0
60	Microtubule-actin cross-linking factor 1	Q9UPN3			N.F.		63	1
61	Zinc finger transcription factor Trps1	Q9UHF7			N.F.		39	1
62	Cdc42 guanine nucleotide exchange factor zizimin 1	Q9BZ29			N.F.		39	1
63	Hypothetical protein KIAA0008.	Q15398			N.F.		12	1
64	COP9 signalosome complex subunit 1	Q13098			N.F.		10	1
65	Hepatocyte growth factor activator precursor	Q04756	✓		N.F.		39	0
66	Conserved oligomeric Golgi complex component 7	P83436			N.F.		12	1
67	Monocarboxylate transporter 1	P53985			N.F.		12	1
68	Antigen KI-67	P46013			N.F.		40	1
69	G2/mitotic-specific cyclin B1	P14635			N.F.		5	1
70	Complement C1s component precursor	P09871	✓	✓	N.F.		27	0
71	Leukocyte common antigen precursor	P08575	✓		N.F.		26	0
72	Apolipoprotein M	O95445			N.F.		6	0
73	Guanine nucleotide exchange factor DBS	O15068			N.F.		26	1
74	Hemoglobin beta chain	P02023	✓		N.F.		2	0
75	*Serum albumin precursor*	*P02768*	✓	✓	*N.A.*		*35*	*23*
76	Ig kappa chain C region	P01834	✓		N.A.		3	3
77	Ig gamma-1 chain C region	P01857		✓	N.A.		9	3
78	Ig gamma-2 chain C region	P01859		✓	N.A.		11	1
79	*Ig alpha-1 chain C region*	*P01876*	✓	✓	*N.A.*		*15*	*2*
80	Ig mu chain C region	P01871	✓		N.A.		12	0
81	Ig gamma-3 chain C region	P01860		✓	N.A.		16	0
82	Ig kappa chain C region	106529		✓	N.A.		3	0
83	Ig alpha-2 chain C region	P01877	✓		N.A.		14	0
84	Ig lambda chain C regions	P01842	✓		N.A.		3	0
85	Ig heavy chain V-III region NIE	P01770			N.A.		2	1
86	Ig delta chain C region	P01880	✓		N.A.		8	0

Table 3. Continued

	Description	Acc. No.	LC^2-2D-PAGE	LC^2-MS/MS	Concentration	# of cysteines	# of non-Cys-peptides
87	Ig heavy chain V-II region NEWM	P01825			N.A.	3	0
88	Ig heavy chain V-I region EU	P01742			N.A.	2	1
89	Immunoglobulin J chain	P01591	✓	✓	N.A.	8	0
90	Ig kappa chain precursor	1082538			N.A.	2	0
91	Ig gamma-4 chain C region	P01861			N.A.	9	0
92	Ig heavy chain V-III region KOL	P01772			N.A.	4	1
93	Immunoglobulin heavy chain VHDJ region	21670047			N.A.	2	0
94	Immunoglobulin kappa light chain variable region	17483730			N.A.	2	0
95	Immunoglobulin heavy chain variable region	11875746			N.A.	2	0
96	Immunoglobulin heavy chain variable region	10636795			N.A.	2	0
97	Immunoglobulin heavy chain variable region	10636671			N.A.	3	0
98	Immunoglobulin gamma heavy chain variable region	10636501			N.A.	3	0
99	Immunoglobulin heavy chain VH-I region	1294810			N.A.	2	0
100	Chain A, Immunoglobin Fc (Igg1) Complexed With Protein G (C2 Fragment)	1065199			N.A.	4	1
101	Anti-herpes simplex virus glycoprotein D Ig heavy chain variable region	1042148			N.A.	2	0
102	Ig Aalpha1 Bur	223099	✓		N.A.	17	0

Plasma proteins that were identified by at least one unique peptide sequence are given by their description and database accession number. The proteins were ranked according to their known concentration in plasma (N.F. indicates that the protein concentration was not found in the literature, N.A. indicates that the protein concentration figures were not applicable since those proteins were (partly) depleted from the plasma sample). The list of identified proteins was screened against the list of proteins identified in a large scale LC/LC-2D-PAGE analysis [26] and in a gel-free LC-LC/MS/MS-analysis [27]. Proteins that were identified in either one of those studies are indicated. The number of cysteine residues in the identified proteins was calculated by the ProtParam-tool which can be found at http://www.expasy.org/cgi-bin/protparam and the number of identified non-cysteine containing peptides is indicated in the last column.

When the enriched plasma sample was digested with trypsin and directly analyzed (*i.e.*, without sorting for cysteine-peptides) by LC-MS/MS only eight, abundant plasma proteins were identified; alpha-1-antitrypsin (P01009), alpha-2-HS-glycoprotein (P02765), alpha-2-macroglobulin (P01023), apolipoprotein A-I (P02647), apolipoprotein A-II (P02652), Ig alpha-1 chain C region (P01876), serum albumin (P02768), and transthyretin (P02766). These proteins were also identified following the developed COFRADIC™ sorting procedure for cysteine-peptides (shown in Table 3 as entries in italics). However, upon simplifying the peptide mixture, 94 other proteins are additionally identified (Table 3), clearly illustrating the power of the developed approach for gel-free proteomics using a 1-D chromatographic system.

4 Discussion

Reversible modification of cysteine residues using Ellman's reagent is an elegant tool to isolate cysteine-containing peptides using COFRADIC™; the reaction is known to be quantitative, the group remains fixed on the cysteines during digestion and RP-HPLC separation and can be easily and selectively removed by a reduction step. Because of the hydrophobic nature of the TNB group, its removal induces a hydrophilic shift which is sufficiently large to enrich cysteine-containing peptides for further gel-free proteome analysis. *In silico* calculations indicate that among the proteomes of different model organisms, about one fifth of all generated and detectable tryptic peptides contain at least one cysteine residue

(data not shown). This implies that isolation of cysteine-containing peptides will lower the complexity of the sample by at least a factor five, which, however, can still be insufficient to analyze every peptide present in the sample and thus some proteins might escape analysis.

In a recently published comprehensive proteome study, a combination of highly-resolving 2-D maps and mass spectrometric analysis was used to identify the products of 123 different open reading frames in a human platelet proteome [25]. Following isolation of cysteine-containing peptides using the COFRADIC™-technology that we developed, we were able to identify 163 different proteins, demonstrating that our technology identifies more proteins and is thus more sensitive than traditional studies. However, during our experiments a weak point of this type of gel-free proteome analysis became clear. Cysteine residues appear to be randomly distributed within protein sequences. From one of our previous studies, we noticed that most protein isoforms differ at their sequence extremities, since by analyzing the COFRADIC™-isolated amino termini of proteins, a clear distinction between many protein isoforms could be made [15]. However, when cysteine-peptides are isolated, the chance is fairly high that the isolated peptides are not spanning the extremities of proteins and therefore, multiple protein isoforms are identified per isolated peptide (see supplementary Table 1). Exactly knowing which protein isoform is present in the analyzed proteome can be important since different isoforms are sometimes organ-specific and may have different biochemical characteristics (*e.g.*, the mammalian sodium pumps, reviewed in [38]). From our experience with traditional proteome studies involving 2-D PAGE, we learned that, while protein isoforms may be separated in 2-D gels, mass spectrometric data generally do not cover sufficiently sequences to specifically identify different isoforms. Indicating that unless monoclonal antibodies specific for protein isoforms are available, most traditional as well as most non-gel proteome studies fail to distinguish protein isoforms.

As can be judged from the list of identified plasma proteins (Table 3), most of these are typical plasma proteins. In order to estimate the dynamic range of our technology, we searched the literature for known concentrations of identified plasma proteins. Among the most abundant, identified plasma proteins are the fibrinogens, serotransferrin and the haptoglobins, which are all present in g/L concentrations. On the other hand, some of the less abundant, known plasma proteins are the complement receptor type 1 protein and calgranulin A which have concentrations in the μg/L range, and sex hormone-binding globulin, coagulation factor V and apolipoprotein C-IV, which have concentration in the low mg/L range (see Table 3). From this data, we can conclude that the developed COFRADIC™ technology for the isolation and identification of cysteine-containing peptides using our Q-TOF mass spectrometer has a dynamic range of 4–5 orders of magnitude, which is consisted with earlier findings [39]. However, the dynamic range of proteins residing in human plasma is much higher and is believed to span at least 10 decades [40]. Therefore, if one would use the developed technology to tackle the human plasma proteome at least 10^5 to 10^6 times more material, *i.e.*, 6.7–67 L of plasma should have been used. Since it was never our intention to identify all the proteins in plasma, such a huge effort has not been made. Clearly, if one would like to do this, LC-based separation techniques, *e.g.*, such as the ones described in [26], could be used prior to COFRADIC™.

Remarkably, more than 400 MS/MS spectra were linked to peptides from human serum albumin (HSA), indicating that the procedure used to deplete serum albumin from plasma was not entirely successful. Other methods that use HSA-specific antibodies for HSA depletion may be more efficient for future analysis and may lead to a better coverage (*e.g.*, [41]). Notwithstanding the limitations of our analysis, more than 100 different proteins have already been identified in a very low volume of plasma (less than 100 μL) (Table 3), a number which can only increase when more starting material is used for these analyses.

In two recent and more comprehensive plasma proteome studies, considerably more material was used for analysis; *e.g.*, 20 mL of serum depleted from all major proteins in the 2-D PAGE study [26] and 500 μL of antibody-depleted serum in the LC-LC/MS/MS study [27]. Hence, in those cases significantly more proteins (about 5 times more) were identified. However, our list contains proteins that were not identified in these two analyses (see Table 3). Among others, these proteins include the known plasma protein apolipoprotein C-IV (P55056), the cytoplasmic protein Cdc42 guanine nucleotide exchange factor zizimin 1 (Q9BZ29), the conserved oligomeric Golgi complex component 7 (P83436), the nuclear antigen KI-67 (P46013) and the nuclear zinc finger transcription factor Trps1 (Q9UHF7). One possible explanation is that the presence of these proteins is the result of destruction of circulating blood cells or endothelial cells. However, if this would happen, the known most abundant household proteins (*e.g.*, those present in circulating platelets (supplementary Table 1)) should be massively present in the list of identified plasma proteins, which is not the case, pointing to the possibility that these proteins might be yet undiscovered endogenous plasma proteins.

As can be seen from Table 3, more than 50% of all identified proteins were identified solely by cysteine-containing peptides, indicating the usefulness of enriching for cysteine-peptides in gel-free proteome studies. Furthermore, the contaminating non-cysteine containing peptides seem to be mainly attributed to the major plasma proteins in the sample (see the entries put in italics in Table 3), which were also identified in a single LC-MS/MS run without cysteine-peptide enrichment (see Section 3). Although such peptides were present in the final mixture, they did not seem to disturb the LC-MS/MS analysis to such a degree that a lot of other peptides were obscured by them. On the contrary, two proteins that do not carry cysteines, a fragment of gamma-fibrinogen and apolipoprotein A-I were actually identified using such peptides.

Finally, it is important to note that the whole COFRADIC™ procedure for isolating cysteine-peptides is largely automated and thus quite fast. Furthermore, we are currently setting up protocols for differential non-gel proteomics using enzymatic incorporation of heavy oxygen atoms at newly formed carboxy termini (*e.g.*, [42]) that are compatible with COFRADIC™-based isolations of representative peptides and have noticed that this type peptide labeling strategy is compatible with all reaction conditions employed during cysteine-COFRADIC™.

K. G. is a Postdoctoral Fellow and L. M. a Research Assistant of the Fund for Scientific Research – Flanders (Belgium) (F.W.O. – Vlaanderen). The project was further supported by the Concerted Research Actions (GOA) of the Flemish Community, the Inter University Attraction Poles (IUAP) and the GBOU-research initiative of the Flanders Institute of Science and Technology (IWT) (GBOU-project No. 20204).

5 References

[1] Washburn, M. P., Wolters, D., Yates, J. R. III., *Nat. Biotechnol.* 2001, *19*, 242–247.

[2] Lipton, M. S., Pasa-Tolic, L., Anderson, G. A., Anderson, D. J. *et al.*, *Proc. Natl. Acad. Sci. USA* 2002, *99*, 11049–11054.

[3] Wu, C. C., MacCoss, M. J., Howell, K. E., Yates, J. R. III., *Nat. Biotechnol.* 2003, *21*, 532–538.

[4] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F. *et al.*, *Nat. Biotechnol.* 1999, *17*, 994–999.

[5] Spahr, C. S., Susin, S. A., Bures, E. J., Robinson, J. H. *et al.*, *Electrophoresis* 2000, *21*, 1635–1650.

[6] Wang, S., Regnier, F. E., *J. Chromatogr. A* 2001, *924*, 345–357.

[7] Shen, M., Guo, L., Wallace, A., Fitzner, J., *et al.*, *Mol. Cell. Proteomics* 2003, *2*, 315–324.

[8] Weinberger, S. R., Viner, R. I., Ho, P., *Electrophoresis* 2002, *23*, 3182–3192.

[9] Ji, J., Chakraborty, A., Geng, M., Zhang, X., *et al.*, *J. Chromatogr. B* 2000, *745*, 197–210.

[10] Oda, Y., Nagasu, T., Chait, B. T., *Nat. Biotechnol.* 2001, *19*, 379–382.

[11] Riggs, L., Sioma, C., Regnier, F. E., *J. Chromatogr. A* 2001, *924*, 359–368.

[12] Zhou, H., Watts, J. D., Aebersold, R., *Nat. Biotechnol.* 2001, *19*, 375–378.

[13] Zhang, H., Li, X. J., Martin, D. B., Aebersold, R., *Nat. Biotechnol.* 2003, *21*, 660–666.

[14] Gevaert, K., Van Damme, J., Goethals, M., Thomas, G. R., *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 896–903.

[15] Gevaert, K., Goethals, M., Martens, L., Van Damme, J., *et al.*, *Nat. Biotechnol.* 2003, *21*, 566–569.

[16] Ellman, G. L., *Arch. Biochem. Biophys.* 1959, *82*, 70–77.

[17] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[18] Sutton, C. W., Pemberton, K. S., Cottrell, J. S., Corbett, J. M., *et al.*, *Electrophoresis* 1995, *16*, 308–316.

[19] Anderson, W. L., Wetlaufer, D. B., *Anal. Biochem.* 1975, *67*, 493–502.

[20] Riddles, P. W., Blakeley, R. L., Zerner, B., *Anal. Biochem.* 1979, *94*, 75–81.

[21] Han, J. C., Han, G. Y., *Anal. Biochem.* 1994, *220*, 5–10.

[22] Wilkins, M. R., Gasteiger, E., Sanchez, J. C., Bairoch, A., Hochstrasser, D. F., *Electrophoresis* 1998, *19*, 1501–1505.

[23] Camon, E., Magrane, M., Barrell, D., Binns, D., *et al.*, *Genome Res.* 2003, *13*, 662–672.

[24] Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., Gygi S. P.,*Proc. Natl. Acad. Sci. USA* 2003, *100*, 6940–6945.

[25] O'Neill, E. E., Brock, C. J., von Kriegsheim, A. F., Pearce, A. C., *et al.*, *Proteomics* 2002, *2*, 288–305.

[26] Pieper, R., Gatlin, C. L., Makusky, A. J., Russo, P. S., *et al.*, *Proteomics* 2003, *3*, 1345–1364.

[27] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J., *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 947–955.

[28] Haeberli, A., *Human Protein Data*, VCH Verlagsgesellschaft mbH, Weinheim 1993.

[29] Available at http://www.fbr.org/publications/plasmaprot/5553POCK.PDF

[30] Kapil, R. P., Axelson, J. E., Mansfield, I. L., Edwards, D. J., *et al.*, *Br. J. Clin. Pharmacol.* 1987, *24*, 781–791.

[31] Yorifuji, H., Anderson, K., Lynch, G. W., Van de Water, L., *et al.*, *Blood* 1988, *72*, 1645–1650.

[32] Crook, M. A., Ch'ng, S. I., Lumb, P., *Blood Coagul. Fibrinolysis* 1999, *10*, 197–200.

[33] Knipping, G., Gogg-Fassolter, G., Frohnwieser, B., Krempler, F., *et al.*, *J. Immunol. Methods* 1997, *202*, 85–95.

[34] Davis, D., Liyou, N., Lockwood, D., Johnson, A., *Clin. Genet.* 2002, *61*, 363–368.

[35] Kotite, L., Zhang, L. H., Yu, Z., Burlingame, A. L., Havel, R. J., *J. Lipid Res.* 2003, *44*, 1387–1394.

[36] Pascal, N., Amouzou, E. K., Sanni, A., Namour, F., *et al.*, *Am. J. Clin. Nutr.* 2002, *76*, 239–244.

[37] Bogumil, T., Rieckmann, P., Kubuschok, B., Felgenhauer, K., Bruck, W., *Neurosci. Lett.* 1998, *247*, 195–197.

[38] Muller-Ehmsen, J., McDonough, A. A., Farley, R. A., Schwinger, R. H., *Basic Res. Cardiol.* 2002, *97*, I25–30.

[39] Clauwaert, K. M., Van Bocxlaer, J. F., Major, H. J., Claereboudt, J. A., *et al.*, *Rapid Commun. Mass Spectrom.* 1999, *13*, 1540–1545.

[40] Anderson, N. L., Anderson, N. G., *Mol. Cell. Proteomics* 2002, *1*, 845–867.

[41] Steel, L. F., Trotter, M. G., Nakajima, P. B., Mattu, T. S., *et al.*, *Mol. Cell. Proteomics* 2003, *2*, 262–270.

[42] Yao, X., Afonso, C., Fenselau, C., *J. Proteome Res.* 2003, *2*, 147.

2.2.8.2. Phospho-COFRADIC

REGULAR ARTICLE

Global phosphoproteome analysis on human HepG2 hepatocytes using reversed-phase diagonal LC

Kris Gevaert, An Staes, Jozef Van Damme, Sara De Groot, Koen Hugelier, Hans Demol, Lennart Martens, Marc Goethals and Joël Vandekerckhove

Department of Medical Protein Research, Flanders Interuniversity Institute for Biotechnology, Ghent University, Ghent, Belgium

We present a phosphoproteomics approach using diagonal RP chromatography as the basic isolation principle. Phosphopeptides present in a tryptic digest of total cellular lysates were first enriched by Fe^{3+}-immobilized metal ion affinity chromatography. Further sorting of the phosphopeptides took place in three steps. First, the resulting peptide mixture was fractionated over reversed-phase chromatography. Second, peptides present in each fraction were treated with phosphatases. Third, the dephosphorylated peptides were then more hydrophobic and shifted towards a later elution interval from the contaminating non-phosphopeptides eluting at the same position as during the primary run. Since the phosphopeptides are isolated as their dephosphorylated form, additional proof for their original phosphorylation state was obtained by split-differential ^{16}O –^{18}O labeling. The method was validated with alpha-casein phosphopeptides and consecutively applied on HepG2 cells. We identified 190 phosphorylated peptides from 152 different proteins. This dataset includes 38 novel protein phosphorylation sites.

Received: October 13, 2004
Accepted: December 27, 2004

Keywords:
Combined fractional diagonal chromatography / MS / Non-gel proteomics / Protein phosphorylation

1 Introduction

In order to obtain a better view on the global regulations of the phosphorylation status of proteins and its biological implications, several holistic attempts have been made to generate qualitative and quantitative data on kinase and phosphatase substrates and to correlate these with normal, altered, or diseased cell stages. This is not an easy task, since a large number of proteins are simultaneously phosphorylated in sub-stoichiometric concentrations and phosphorylation can occur on different residues such as serine, threonine, and tyrosine next to histidine [1], lysine [2], and arginine [3] – modifications that all have specific chemical characteristics and stabilities.

As early as 1975, Kang and co-workers [4] used 2-D gel analysis of ^{32}P-labeled proteins for a global differential analysis. However, it took ten more years to develop a method able to identify gel-separated proteins by micro-sequencing [5] and later by fast PMF [6]. While it is routine to identify phosphorylated proteins, locating the actual phosphorylation sites still remains challenging [7]. In addition, 2-D gel-based phosphoproteomics encompasses problems inherently linked to limitations of the gel approach, such as limited sensitivity, loss of proteins with extreme properties, *etc.* Recently, alternative strategies were developed by different groups. Central in many approaches is a highly resolving chromatographic system, or combinations thereof, combined with high-sensitivity MS for peptide identification [8–12]. Other approaches rely on chemical modification of the

Correspondence: Professor Kris Gevaert, Department of Medical Protein Research, Flanders Interuniversity Institute for Biotechnology, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium
E-mail: kris.gevaert@ugent.be
Fax: +32-92-649484

Abbreviations: AU, absorbance units; **COFRADIC**, combined fractional diagonal chromatography

phosphorylation sites creating derivatives such as phosphoramidates [13], biotinylated amino acids [14, 15], and thiols [16], thereby generating sites that can be used in affinity-adsorption strategies.

We recently developed a non-gel technology called combined fractional diagonal chromatography (COFRADIC) to isolate sets of representative peptides from digests of complete proteomes. The basic principle of COFRADIC is to induce a chromatographic shift by modifying a specific subset of peptides. This alteration is done between two consecutive, identical separations. In the second run, altered peptides elute differently from their original position and segregate from the bulk of unaltered peptides that elute unchanged.

So far, COFRADIC was applied for the isolation of methionyl-peptides [17], cysteinyl-peptides [18], and peptides covering the amino termini of proteins [19]. We have now modified COFRADIC to isolate phosphorylated peptides in trypsin digests of cell lysates. As for the previous systems, we used a chromatographic shift, after phosphatase treatment of peptide fractions collected during the primary run. In the secondary run, ex-phosphopeptides shift towards later elution times and are identified by MS/MS analysis. Here we provide details and an initial assessment of the procedure on a total lysate of HepG2 cells.

2 Materials and methods

2.1 Sorting and identification of phosphorylated α-S1 casein peptides

In total, 1 nmol of bovine α-casein (70% pure; Sigma-Aldrich, St. Louis, MO, USA) was digested for 4 h with 1 μg of trypsin (sequencing grade, modified porcine trypsin from Promega Corporation, Madison, WI, USA) in 100 μL freshly prepared 50 mM ammonium bicarbonate. The generated peptide mixture was separated by RP-HPLC onto a RP-HPLC column (2.1 mm id × 150 mm 300SB-C18 column, ZorbaxR; Agilent Technologies, Waldbronn, Germany) using an Agilent 1100 Series HPLC system. The column was first washed for 10 min with solvent A (10 mM ammonium acetate at pH 5.5 in ACN/water (2/98, v/v), both Baker HPLC analyzed; Mallinckrodt Baker, Deventer, The Netherlands), and a linear gradient to 100% solvent B (10 mM ammonium acetate at pH 5.5 in ACN/water (70/30, v/v)) was applied over 100 min. Using Agilent's electronic flow controller box, a constant flow of 80 μL/min was generated.

This separation is referred to as the primary run. From 40 min onwards until 80 min, fractions of 2 min each (160 μL) were collected in a microtiter plate. From each primary fraction, 5 μL was spotted on a MTP AnchorChip™ 600/384 (Bruker Daltonics, Bremen, Germany) that was precoated with a thin layer of MALDI-matrix (0.25 mg of CHCA dissolved in 1 mL of acetone). The samples were automatically analyzed in reflectron mode with a Bruker Ultraflex TOF/TOF mass spectrometer using the AutoXecute module of Bruker's flexControl 2.2. for each sample, 300 shots were acquired.

This fast mass screening allowed us to readily identify fractions that might contain putative α-casein phosphopeptides. These fractions were retained, dried, and re-dissolved in 50 μL of calf intestinal alkaline phosphatase (CIP) reaction buffer as delivered by the manufacturer (New England Biolabs, Beverly, MA, USA) and a total of 10 U of CIP was used *per* fraction to carry out the dephosphorylation. After 1 h at 37°C, 50 μL of RP solvent A was added to the reaction mixture and each of the fractions of interest was re-separated on the same RP column. This is called the secondary run. Because we were expecting a hydrophobic shift associated with dephosphorylation, we collected the peptides eluting in a 20 min interval, following the position of the non-altered peptides of which the elution time was not affected due to identical separation conditions. Material eluting in each delayed zone was collected in 8 μL droplets on a Bruker Anchor Chip™ and analyzed as described.

2.2 Cell lysis and peptide pretreatments

HepG2 cells were grown at 37°C in DMEM complemented with 10% v/v heat-inactivated fetal calf serum, 25 U/mL of penicillin, and 25 μg/mL of streptomycin as previously described [20]. Cells were grown to sub-confluent levels and stimulated with 10 μM forskolin (Sigma-Aldrich) for 30 min at 37°C. The cells were detached in enzyme-free PBS-based cell dissociation buffer (Invitrogen, Carlsbad, CA, USA) for 15 min at 37°C, briefly centrifuged, and extensively washed in PBS. A pellet containing 25 million HepG2 cells was dissolved in 5 mL of 4 M urea and 0.625% w/v CHAPS in 100 mM NaH_2PO_4 (pH 7.4). The cell disruption buffer further contained protease inhibitors (Complete EDTA-free protease inhibitor cocktail tablets; F. Hoffmann-La Roche, Basel, Switzerland), phosphatase inhibitors (10 mM NaF, 200 μM sodium orthovanadate, 20 mM β-glycerophosphate, 5 μM phenylvalerate, and 2 mM levamisole hydrochloride) as well as 10 mM Tris(2-carboxyethyl)phosphine (TCEP) and 100 mM iodoacetamide. Cell disruption started at room temperature for 5 min and continued on ice for another 30 min. During this procedure, the cells were kept in the dark. Disrupted cells were briefly sonicated and centrifuged for 10 min at 13 000 × *g* at 4°C and the pellets were discarded.

The protein solution was desalted onto two PD-10 columns (Amersham Biosciences, Uppsala, Sweden) and collected in a total of 7 mL of protein digestion buffer consisting of 2 M of fresh urea in 100 mM Tris HCl (pH 8.7) and phosphatase inhibitors (see previous paragraph). The lysate was incubated overnight at 37°C with trypsin at an enzyme/substrate ratio of 1/100 w/w. A light precipitate that formed during the digestion was removed by centrifugation at 13 000 × *g* after adjusting the pH to 3 with 1 M HCl. Phosphopeptides in this mixture were enriched on Fe^{3+}-loaded IMAC beads (Sigma-Aldrich) at pH 3.0 according to the

manufacturer's protocol. Desorption from the IMAC resin was done in a total volume of 2 mL of 0.4 M ammonium hydroxide. This mixture was divided into two equal parts and dried in a centrifugal vacuum concentrator.

One part was re-dissolved in 50 μL of CIP reaction buffer as delivered by the manufacturer (New England Biolabs) and a total of 10 U of CIP, 200 U of lambda protein phosphatase (Upstate, Milton Keynes, UK), and 0.66 U of alkaline *Escherichia coli* phosphatase (Sigma-Aldrich) were added to dephosphorylate all phosphopeptides. This reaction proceeded for 1 h at 37°C after which the pH was lowered to pH 5 by adding 80 μL of 0.5 M KH_2PO_4. Subsequently, 4 μg of trypsin was added and the mixture was dried to complete dryness. It was then re-dissolved in 150 μL $H_2{}^{18}O$ (93.7% w/w pure; ARC Laboratories, Amsterdam, The Netherlands). The trypsin-mediated oxygen exchange was continued overnight at 37°C. Finally the $H_2{}^{18}O$-peptide solution was transferred to an Eppendorf tube in which 1.5 μmol of TCEP and 15 μmol of iodoacetamide were present as a dried pellet. Trypsin inactivation by reductive alkylation proceeded at pH 8.5 and resulted in the inhibition of the $^{18}O/^{16}O$ back-exchange during further steps of the procedure [21].

The second part of the IMAC-enriched peptide mixture was subjected to the same procedure except for addition of phosphatases and replacement of $H_2{}^{18}O$ by natural water.

2.3 COFRADIC isolation of ex-phosphopeptides from total lysate

Following trypsin inactivation, the two peptide mixtures were mixed and loaded on an Agilent 1100 Series HPLC system run under the conditions described in Section 2.1. Peptides eluting between 22 and 70 min were collected in 48 fractions of 1 min each. Primary fractions that were separated by 16 min (*e.g.*, fractions 5, 21, and 37 or fractions 6, 22, and 38 *etc.*) were pooled, dried, and reconstituted in 50 μL of buffer containing the phosphatase cocktail (see Section 2.2). The dephosphorylation reaction proceeded for 1 h at 37°C and was terminated by lowering the pH to 5.5 after adding 40 μL of solvent A and 10 μL of 50 mM acetic acid. The pooled and dephosphorylated peptide fractions were re-separated onto the same column system and under identical chromatographic conditions.

Phosphopeptides which were dephosphorylated as the result of the phosphatase treatment between the primary and secondary run displayed a hydrophobic shift compared to their original elution time. A second characteristic of the peptides was the absence of the ^{18}O-label such that they consisted of the ^{16}O-isotopic envelope only. They were collected over 14 min. This time interval started 1 min after each elution peak of the non-modified peptides. For the examples given above, this means that the ex-phosphopeptides of primary fraction 5 eluted between 6 and 20 min, fraction 21, between 22 and 36 min, and fraction 37, between 38 and 52 min. This procedure yielded 48 secondary fractions of 1.12 mL each. These were dried and re-dissolved in 20 μL of 0.05% formic acid in 2/98 v/v ACN/water (solvent A). Half of this solution was loaded on a trapping column (0.3 mm id × 5 mm, PepMap; LC Packings, Amsterdam, The Netherlands) at a flow rate of 20 μL/min of solvent A for 5 min using a CapLC® system (Waters Corporation, Manchester, UK). The trapping column was back-flushed, whereby the sample was loaded on a nano-scale RP C18 column (75 μM id × 150 mm PepMap column; LC Packings) and peptides were eluted from this column using a gradient from 0 to 100% solvent B (0.05% formic acid in 70/30 ACN/water) at 200 nL/min applied over 50 min. The outlet of the nano-column was connected with a distal metal-coated fused silica PicoTip™ needle (PicoTip™ FS360-20-10-D-C7; New Objective, Woburn, MA, USA) and placed in front of the inlet of a Q-TOF1 mass spectrometer (Micromass UK, Cheshire, UK). Data-dependent acquisition began 20 min after the solvent gradient was started and was such that doubly and triply charged ions were selected for fragmentation. Following each LC-MS/MS run, the obtained fragmentation spectra were automatically converted to Micromass proprietary pkl format using ProteinLynx from the Micromass Mass Lynx software (version 3.4).

2.4 Identification of peptides by MASCOT and evaluation of their *in vivo* phosphorylation status

All the obtained pkl files were merged in blocks of 299 *per* MASCOT search (MASCOT version 1.9). MASCOT Daemon (version 1.9) was used in the follow-up search setting, in such a way that first the Swiss-Prot database (downloaded at ftp://us.expasy.org/databases/swiss-prot) was searched (with restriction to human proteins) and those pkl files that were not identified were subsequently searched against all human protein sequences present in NCBI's non-redundant protein database (downloaded at ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/protein). MASCOT search parameters were set as follows: enzyme, trypsin; maximum number of missed cleavages, 1; fixed modification, carbamidomethyl (C); variable modifications, deamidation (NQ), oxidation (M), *N*-acetyl (protein), and pyroglutamate formation (*N*-terminal Q); peptide mass tolerance, ±0.3 Da; MS/MS tolerance, ±0.3 Da and instrument setting was set at ESI-QUAD-TOF. Only MS/MS spectra that exceeded the MASCOT identity threshold at the 95% significance level were withheld for further analysis. Such spectra were examined for the presence of typical peptide fragment ions (b- and y-type of fragment ions) and only when the MASCOT scores were substantially higher than the corresponding identity threshold score and sufficient peptide sequence coverage was observed (typically over 50%) were these identifications considered positive.

Once identified by MASCOT, the mass spectrum (obtained in MS mode) from the peptide was evaluated to check whether ^{18}O-isotopes were incorporated into the peptide's carboxyl end group (false positive) or not (ex-phosphorylated peptide). Only those peptides that did not contain the heavy oxygen isotopes were withheld as peptides that were derived from *in vivo* phosphorylated and are listed in Supplementary Table 1.

Table 1. List of identified phosphorylation sites in HepG2 proteins. Phosphorylated peptides that were identified following targeted MS/MS analysis of the primary COFRADIC fractions are given in bold. Sorted peptides containing only one possible phosphorylation site (serine, threonine, or tyrosine) are given in normal type. The identified phosphorylation sites are underlined and indicated in the penultimate column. Literature links to previous studies in which these phosphorylation sites were identified are given in the last column

	Acc. no.	Start	End	Identified peptide	Protein description	Phosphorylation site	Reference
	Phosphoserine-containing peptides						
1	P05386	98	113	**KEESEESDDDMGFGLF**	60S acidic ribosomal protein P1	S-104	[10, 12]
		98	114	**KEESEESDDDMGFGLFD**	S-104		[10, 12]
2	P12956	218	229	DIISIAEDEDLR	ATP-dependent DNA helicase II, 70 kDa subunit	S-222	
3	Q14019	111	117	EFVISDR	Coactosin-like protein	S-115	
4	P17812	571	584	**SGSSSPDSEITELK**	CTP synthase	S-571, 573 & 574	[11]
5	P29692	144	169	**KPATPAEDDEDDDIDLFGS DNEEEDK**	Elongation factor 1-delta	S-162	
6	Q99613	34	47	**QPLLLSEDEEDTKR**	Eukaryotic translation initiation factor 3 subunit 8	S-39	
7	Q08379	7	14	QSKLAAAK	Golgi autoantigen, golgin subfamily A member 2	S-8	
8	P04792	80	89	**QLSSGVSEIR**	Heat shock 27 kDa protein	S-82	[27]
9	13375676	137	151	IGELGAPEVWGLSPK	Hypothetical protein FLJ22626	S-149	
10	Q9Y4I1	1209	1217	QELESENKK	Myosin Va	S-1213	
11	Q9H1E3	105	128	EMLMEDVGSEEEQEEE DEAPFQEK	Nuclear ubiquitous casein and cyclin-dependent kinases substrate	S-113	
12	Q15149	773	784	EEEEVGFDWSDR	Plectin 1	S-782	
13	P51531	1559	1575	**AKPVVSDFDSDEEQDER**	Possible global transcription activator SNF2L2	S-1564 & 1568	
14	Q96B49	61	74	NLSDIDLMAPQPGV	Similar to over-expressed breast tumor protein	S-63	
15	P26368	2	9	**SDFDEFER**	Splicing factor U2AF 65-kDa subunit	S-2	
16	P35269	207	230	**IHDLEDDLEMSSDASDAS GEEGGR**	Transcription initiation factor IIF, alpha subunit	S-217, 218, 221 & 224	
17	Q13595	2	12	**S**DVEENNFEGR	Transformer-2 protein homolog	S-2	
18	Q13144	542	551	GGSPQMDDIK	Translation initiation factor eIF-2B epsilon subunit	S-544	
19	O60841	131	145	**VEMYSGSDDDDDFNK**	Translation initiation factor IF-2	S-135 & 137	
		131	148	**VEMYSGSDDDDDFNKLPK**		S-135 & 137	
	Phosphothreonine-containing peptides						
1	P12814	48	60	AGTQIENIEEDFR	Alpha-actinin 1	T-50	
2	Q9NY33	424	432	LTFLEEDDK	Dipeptidyl-peptidase III	T-425	
3	P04075	60	68	QLLLTADDR	Fructose-bisphosphate aldolase A	T-65	
4	P09651	130	139	IEVIEIMTDR	Heterogeneous nuclear ribonucleoprotein A1	T-138	
5	Q15047	836	844	ILTDDFADK	Histone-lysine *N*-methyltransferase, H3 lysine-9 specific 4	T-838	
6	Q01650	499	507	LMQVVPQET	Large neutral amino acids transporter small subunit 1	T-507	
7	Q96PK2	5624	5635	LEMTAVADIFDR	Microtubule-actin crosslinking factor 1, isoform 4	T-5628	
8	29743831	151	160	KITIADCGQL	Similar to peptidyl-Pro *cis trans* isomerase	T-153	
9	P10599	96	104	LEATINELV	Thioredoxin	T-100	
10	P18206	502	511	WIDNPTVDDR	Vinculin	T-507	
11	P12955	188	196	TDMELEVLR	Xaa-Pro dipeptidase	T-188	

Table 1. Continued

	Acc. no.	Start	End	Identified peptide	Protein description	Phosphorylation site	Reference
	Phosphotyrosine-containing peptides						
1	Q99460	45	55	IEVL**Y**EDEGFR	26S proteasome non-ATPase regulatory subunit 1	Y-49	
2	O00571	342	350	**Y**LVLDEADR	DEAD-box protein 3	Y-342	
3	P13639	265	272	**Y**FDPANGK	Elongation factor 2	Y-265	
4	P00533	978	986	**Y**LVIQGDER	Epidermal growth factor receptor precursor	Y-978	
5	P30501	75	86	WVEQEGPE**Y**WDR	HLA class I histocompatibility antigen, Cw-2 alpha chain precursor	Y-83	

3 Results

3.1 The phosphorylation sites of α-S1 casein: A feasibility study

This experiment was undertaken to study the possibility of using the diagonal chromatography approach to isolate and identify phosphorylation sites in phoshoproteins. At this stage it was important to learn about the extent and reproducibility of the hydrophobic shift during RP chromatography associated with the removal of one or more phosphate groups from the phosphopeptides.

Therefore we used a tryptic digest of bovine α-casein that was fractionated by RP-HPLC (Fig. 1A). The method is illustrated with the peptide mixture present in the fraction eluting between 46 and 48 min (Fig. 1A), containing a peptide with a mass of 1952.00 Da $(M+H)^+$ indicative of an α-S1-casein tryptic peptide with the expected phosphorylation site at serine-130 (^{119}YKVPQLEIVPNpSAEER^{134}). The MALDI-MS spectrum of the components of this fraction is shown in Fig. 1B. After phosphatase treatment, the secondary run (Fig. 1C) was analyzed in 8 μL aliquots (see Section 2) in an elution interval between 48 and 68 min. Nearly all previously identified peptides were observed in the non-shifted area (Fig. 1D), while the only prominent peptide that was missing was now found in its dephosphorylated form (1871.99 Da) shifted by about 2 min (Fig. 1C and E). Its identity was confirmed by fragmentation analysis (Fig. 1F).

The UV profile in the 48 to 68 min interval shows two additional peaks which could not be assigned to peptides. Most likely these are contaminants introduced from the CIP enzyme preparation. Peaks in front of the selected fraction similarly represent UV-absorbing material mixed with a small amount of CIP degradation products (Fig. 1C).

A repetition of this analysis on other fractions of interest led to the assignment of phosphorylation sites at serines 61 and 63 (results not shown). Based on these results, we could now set the parameters for a global phosphoproteome study.

3.2 Parameters for a global phosphoproteome analysis

Although the diagonal chromatography principle was already successful, there were further additional steps necessary in order to tackle a global analysis on total cellular lysates. First, in contrast to the situation for α-casein, phosphopeptides from cell lysates seldom represent major components. Therefore we decided to enrich for phosphopeptides using Fe^{3+}-IMAC beads. The retained peptide mixture still contained a high number of non-phosphorylated (mainly acidic) peptides. However, the latter could be easily separated from the phosphopeptides by the phosphatase-induced hydrophobic shift as shown in Section 3.1. Second, the extent of the hydrophobic shift depends on the nature of the peptide, the nature of the phosphorylated amino acid, the number of phosphate groups removed, and the chromatographic system used. Shifts observed in the ammonium acetate (pH 5.5)/ACN system were generally larger than those in the 0.1% TFA/ACN system. The former was therefore selected for further experiments. Here loss of one phosphate group was typically associated with an average delay of 3 min but was greater when phosphotyrosine was present. Shifts of poly-phosphorylated peptides are generally much larger and more difficult to predict. Therefore, we decided to use an interval of 14 min to collect the shifted, dephosphorylated peptides. Third, in our approach phosphopeptides are recovered in their dephosphorylated form (as ex-phosphopeptides). Thus, at this stage there is no trace left over from the previous phosphorylation status (such as a chemical tag or a heavy isotope). This creates a serious risk of occasional capture of normal peptides shifting due to slight variations in the chromatographic system and leading to an incorrect recording as ex-phosphopeptides. In order to solve this problem, we introduced a split-differential labeling approach. Prior to the sorting procedure, the tryptic peptide mixture is divided into two equal parts. In one part peptides are dephosphorylated and also stably tagged with two ^{18}O-atoms at the COOH-termini. The second part is

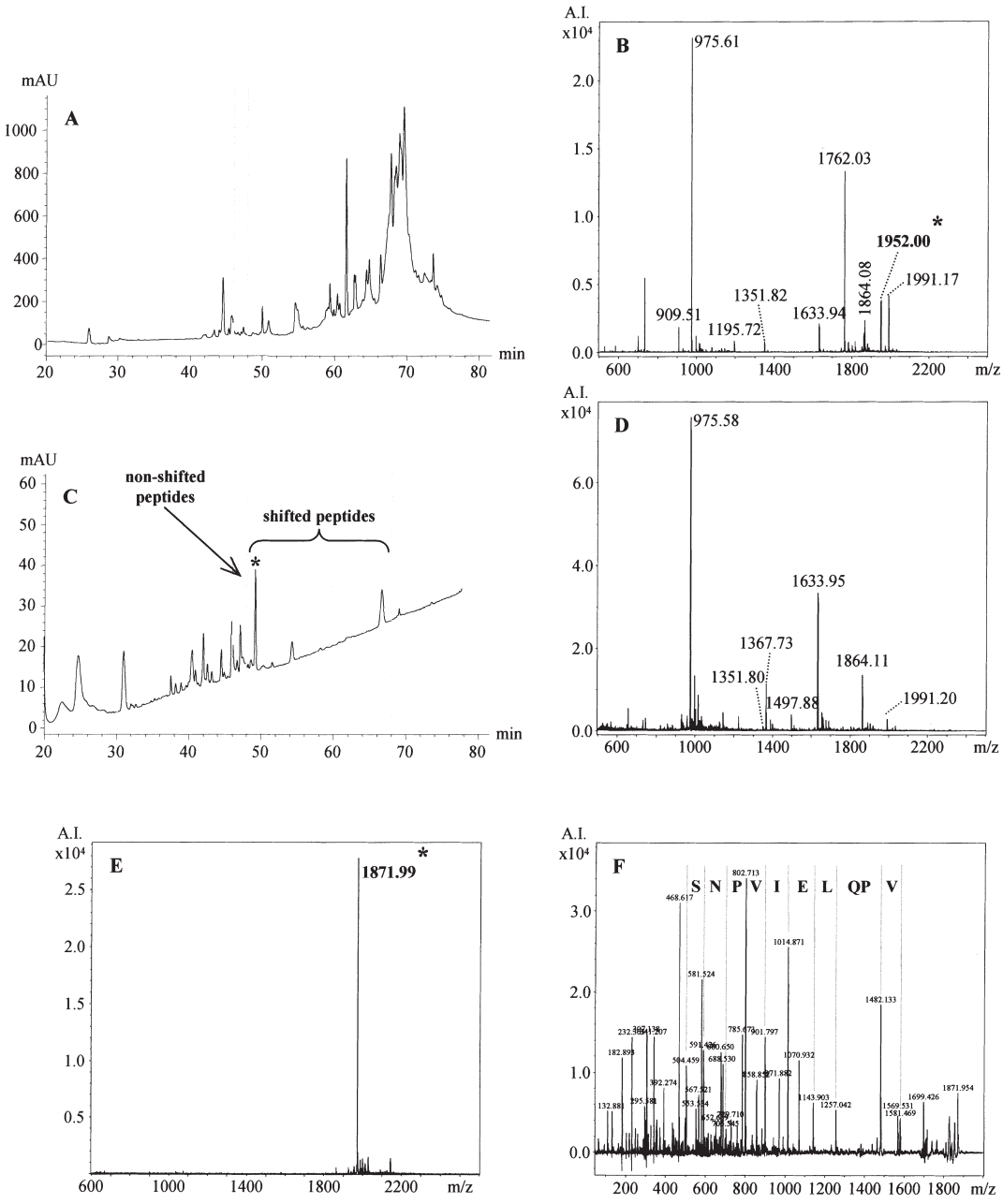


Figure 1. Sorting of a phosphorylated tryptic peptide from bovine α-S1 casein. (A) UV absorption chromatogram (214 nm) of 1 nmol of a tryptic digest of bovine α-casein. (B) MALDI mass spectrum of the primary fraction delineated in (A); a phosphorylated peptide carrying phosphoserine-130 of α-S1 casein is indicated with an asterisk. (C) Secondary RP-HPLC separation (UV absorption chromatogram at 214 nm is shown) of the selected primary fraction that was treated with phosphatases. (D, E) MALDI mass spectra of the non-shifted (primary collection interval) and shifted, ex-phosphorylated peptides (indicated with an asterisk in (C)). (F) MALDI-PSD spectrum of the ex-phosphorylated peptide shown in (E); the sequence tag that could be derived using the y-type of fragment ions is indicated and led to the validation of the peptide sequence ^{119}YKVPQLEIVPNSAEER^{134} of α-S1 casein.

neither dephosphorylated nor labeled with ^{18}O-isotopes. Then both parts are mixed again for further sorting. At this stage, the peptide mixture is composed of three types of peptides: (a) peptides that were never phosphorylated *in vivo,* present as ^{16}O/^{18}O-couples differing by 4 amu in a 1/1 ratio, (b) peptides that were dephosphorylated by the phosphatase treatment of the part that carries the ^{18}O-isotopes, and (c) their phosphorylated counterparts derived from the second part and carrying only the natural oxygen isotopes. During the primary run there is already a separation of the

^{16}O-labeled phosphopeptides from their dephosphorylated counterparts that are ^{18}O-labeled in as much that they elute in different primary fractions. The consecutive phosphatase treatment taking place between the primary and secondary run now only affects the ^{16}O-labeled phosphopeptides, because their ^{18}O-labeled variants were dephosphorylated prior to the first run. As a consequence of this split-differential labeling, only the phosphorylated peptides labeled with the ^{16}O-tag will show shifts. Other non-phosphorylated peptides which might be caught accidentally in the shifted zone will appear as ^{16}O/^{18}O-doublets and can be distinguished from the real ex-phosphopeptides. Fourth, in order to reach a broad specificity towards all types of phosphorylation sites, we now used a cocktail of phosphatases including the phage lambda phosphatase and the *E. coli* alkaline phosphatase in addition to the CIP used in the α-casein experiment. A general layout of the different steps used in the improved protocol is represented in Fig. 2. For technical details, refer to Section 2.3.

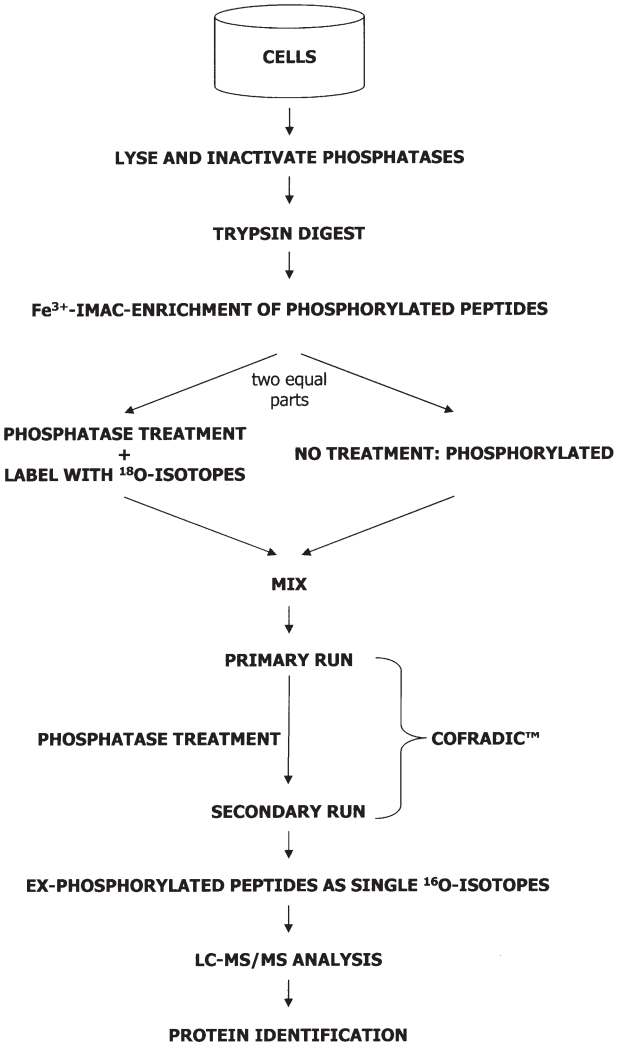


Figure 2. Scheme of the procedure leading to the isolation of phosphorylated peptides and their localization in phosphorylated proteins.

3.3 The phosphoproteome of forskolin-stimulated cultured HepG2 cells

A total of 25 million HepG2 cells were used for phosphopeptide isolation. Following COFRADIC MS/MS analysis of the isolated peptides and data analysis, 294 spectra were from ex-phosphorylated peptides and these were linked to 190 peptides (Supplementary Table 1). All together these peptides identified 152 different phosphoproteins in HepG2 cells. Figure 3 shows a typical MS spectrum of a secondary sub-fraction. Insets show the isotope envelopes of the ^{16}O-peptides assessed as ex-phosphorylated peptides. One ^{16}O/^{18}O-doublet is shown (m/z = 670.78 Th) which is considered to be a false positive.

An important aspect of our approach is the isolation of peptides in their dephosphorylated state. In this respect, two questions are eminent: are we really dealing with ex-phosphopeptides, and when there are multiple potential phosphorylation sites, which sites are phosphorylated *in vivo*? The most direct answer is to isolate and sequence peptides in their phosphorylated forms. Knowing that the phosphorylated derivatives of the peptides listed in Supplementary Table 1 should be searched for in the primary fractions as ^{16}O-singlets carrying additional masses of (multiples of) 80 Da, it is possible to start a targeted analysis in each primary fraction. Since these fractions are much more crowded than the secondary fractions containing the sorted peptides, and since phosphopeptides are not easily ionized in acidic buffers in a background of unphosphorylated peptides, we could only directly assign a limited number of phosphorylation sites (Table 1). The assignment of serine-39 as a previously unnoticed phosphorylation site in the eukaryotic translation initiation factor 3, subunit 8 is documented in Fig. 4. Following COFRADIC, the peptide bearing this phosphorylation site was identified as ^{34}QPLLLSEDEEDTKR^{47}, starting with a pyroglutamic acid and containing two possible phosphorylation sites: serine-39 and threonine-45 (Fig. 4A). A mass inclusion list using the observed mass of the ex-phosphorylated peptide (1655.86 Da, $(M+H)^+$) was generated, including the possibility that either one (+80 Da) or both sites (+160 Da) were phosphorylated and that these peptides are ionized to their doubly or triply charged form. When the primary fraction from which this peptide was isolated was analyzed in LC-MS/MS mode with this mass inclusion list, a doubly charged peptide with an m/z value of 868.36 Th was fragmented. A neutral loss of 98 Da is evident in the obtained MS/MS spectrum, while a loss of 80 Da is not observed (Fig. 4B). This already hints to a phosphorylated serine or threonine residue present in the analyzed peptide [22]. Upon comparison of the MS/MS spectra of the dephosphorylated and the phosphorylated peptide and closer inspection of the b-type of peptide fragments (b_5 to b'_8), the serine residue in the peptide was found to be converted to dehydroalanine, which is consistent with beta-elimination of phosphorylated serines observed in MS/MS spectra [23]. Furthermore, the y-type of fragment ions indicated that ser-

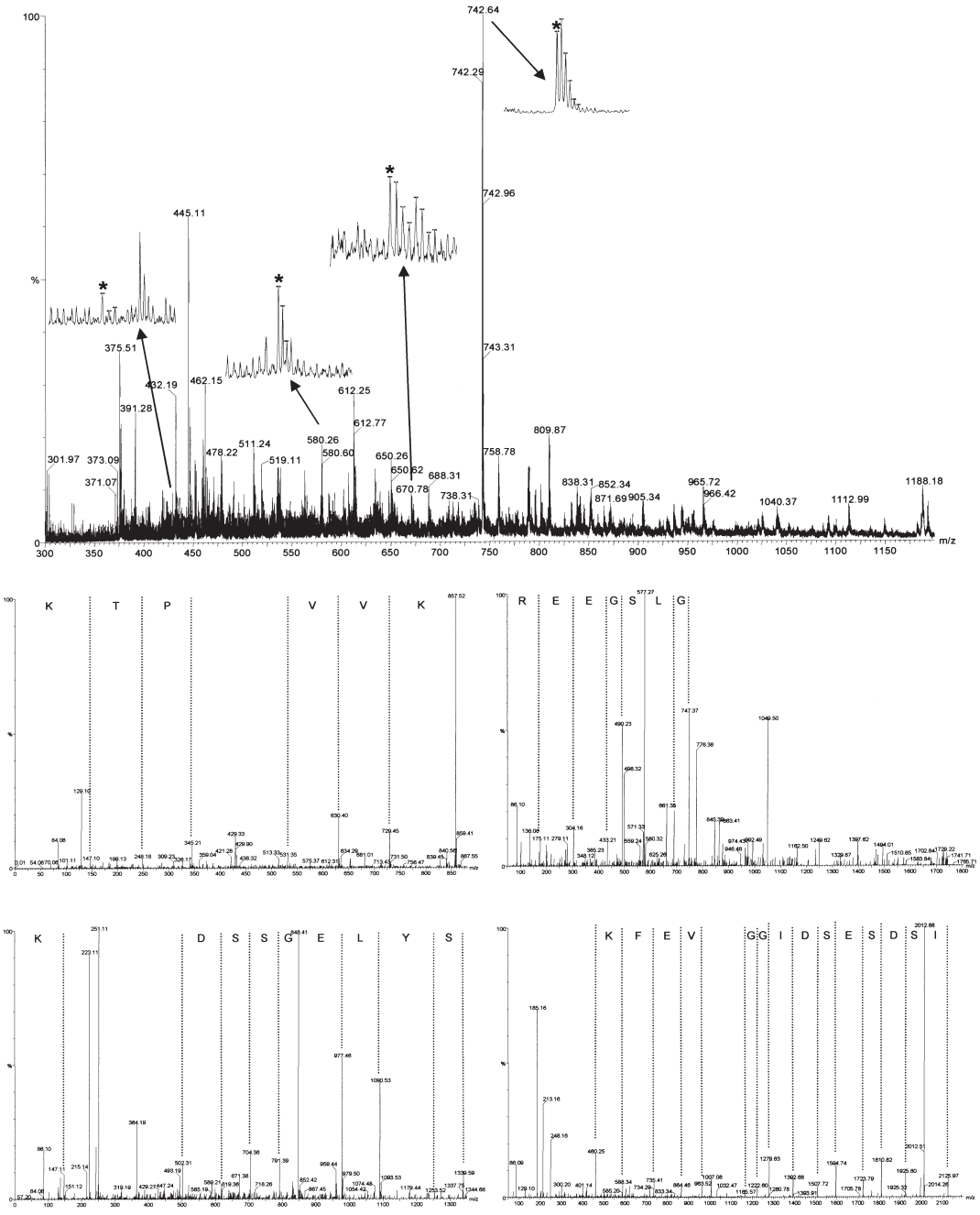


Figure 3. Identification of phosphopeptides following MS/MS analysis of their ex-phosphorylated counterparts. A typical mass spectrum obtained during LC-MS analysis of a secondary COFRADIC fraction is shown in the upper panel. The insets show zoomed regions of this spectrum and the ions indicated with an asterisk were chosen for MS/MS analysis. The isotope envelopes of the ions with *m*/*z* values of 429.25 Th (doubly charged), 580.26 Th (triply charged), and 742.29 Th (triply charged) do not contain any incorporated ^{18}O-isotopes and thus belong to sorted ex-phosphorylated peptides. Following MS/MS analysis of these ions, the following peptides were identified: ^{63}KVVVSPTK^{70} (429.25 Th) from nucleolin (left spectrum of middle panel), ^{759}REVLYDSEGLSGEER^{773} from the FLJ00034 protein (right spectrum of middle panel), and ^{250}VISDSESDIGGSDVEFKPDTK^{270} from the DNA mismatch repair protein MSH6 (right spectrum of lower panel). As can be seen in Supplementary Table 1, all these peptides are predicted to contain kinase substrates. The ion with a *m*/*z* value of 670.78 Th (triply charged) contains ^{18}O-isotopes (see inset). It could either be a false positive or a peptide which was not fully dephosphorylated before the first run (see text). This peptide was also selected for MS/MS analysis (left spectrum of lower panel) and the sequence ^{672}SYLEGSSDNQLK^{683} from the DEAD-box protein 20 (G.I. number 12643886) was identified (in all MS/MS spectra shown, only the y-type of fragment ions and thereby the visualized sequence tags are indicated).

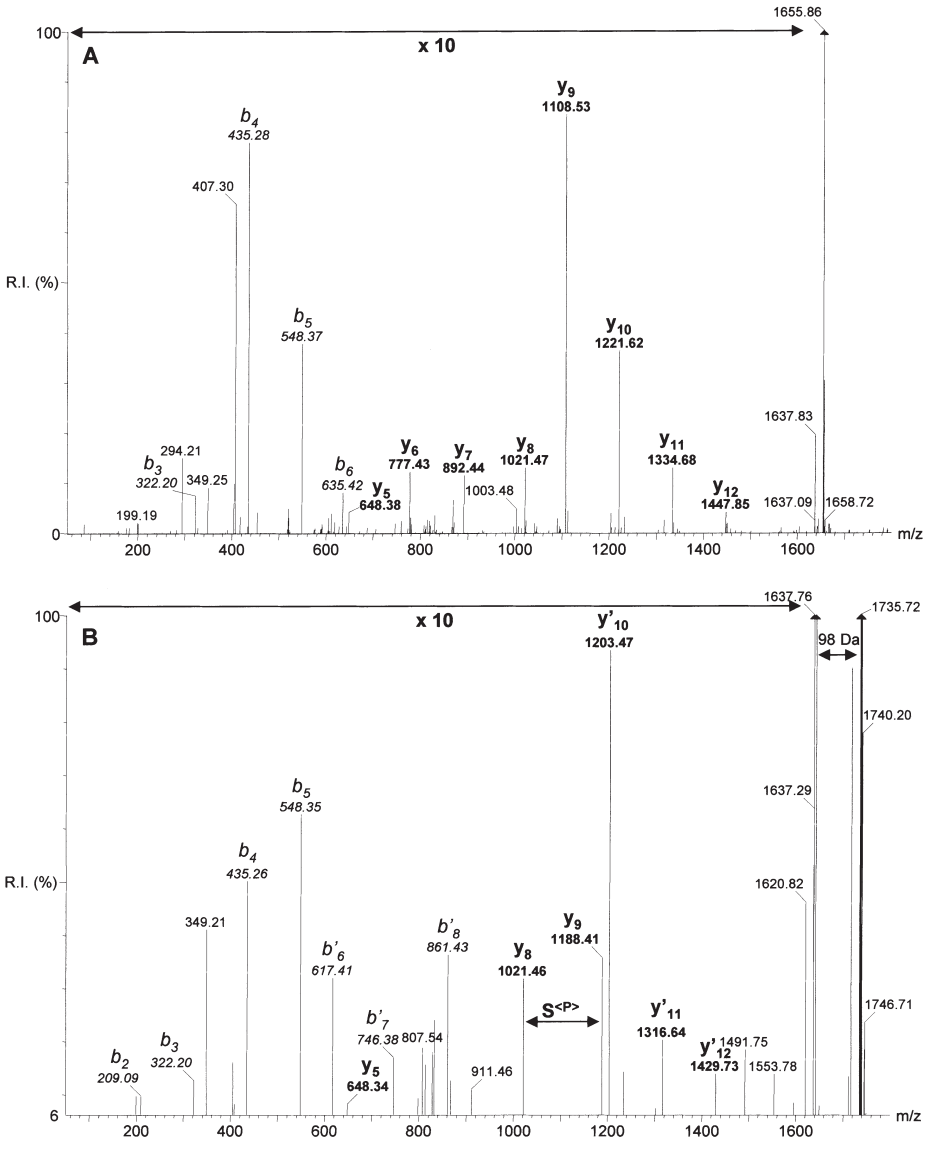


Figure 4. Characterization of serine-39 as an *in vivo* phosphorylation site of the eukaryotic translation initiation factor 3, subunit 8. Following COFRADIC isolation, a peptide with the sequence ^{34}QPLLLSEDEEDTKR^{47} (first residue is a pyroglutamic acid) was identified; the observed b- and y-type peptide fragments are indicated (A). Using a mass inclusion list, the phosphorylated counterpart of this peptide could be analyzed in MS/MS mode in the corresponding primary fraction (B). Upon comparison of the two fragmentation spectra and closer inspection of the observed fragment ions, the serine residue was found to be phosphorylated *in vivo*. In particular, a mass difference of 166.95 Da was found between the y_8 and y_9 ion, corresponding to the mass of the phosphorylated serine ($S^{<P>}$). Upon CAD, phosphoserine is frequently converted to dehydroalanine as evident from the massive loss of 98 Da of the precursor ion and the difference of 18 Da between the predicted b and y fragments and the ones observed (indicated by hyphens in panel B).

ine-39 was phosphorylated; a mass difference of 166.95 Da was observed between the y_8 and y_9 ion. Taken together these results allude to a possible control of the activity of this particular translation initiation factor by phosphorylation of serine-39.

The list of identified peptides further contains 26 peptides that carry only one potential phosphorylation site. Given the experimental route that these peptides followed, one can state that for these peptides the *in vivo* phosphorylated amino acid is hereby identified. Combining this sub-list with the list obtained after targeted analysis of phosphorylated peptides present in the primary fractions leads to the identification of 38 novel phosphorylation sites in 33 different human proteins (see Table 1, without taking into account the known phosphorylation sites of CTP synthase, hsp27, and the 60S acidic ribosomal protein).

4 Discussion

Our approach is selective for phosphopeptides at two levels: we enriched by Fe^{3+}-IMAC and we induced a shift by dephosphorylation during diagonal chromatography.

The use of IMAC has been debated, since it was reported to display different degrees of specificity depending on the transition element used [9, 34]. In addition, IMAC appears biased toward peptides carrying multiple carboxyl groups. Here, we did not try other metal ions, neither did we make methylesters to improve the specificity [8]. Rather, we relied on the second step: selecting on the susceptibility of phosphopeptides for phosphatases. Our procedure partly resembles that described by Bonenfant *et al.* [35]. In the latter, differential protein phosphorylation is measured by a three-step process including stable isotope labeling, IMAC enrichment,

and alkaline phosphatase treatment to release IMAC-retained peptides. While this procedure is suited for the analysis of simple protein mixtures, our procedure is more adapted for the analysis of complex mixtures, because of the diagonal chromatography step which provides an additional but necessary peptide selection.

The method described here will only reach its full potential when IMAC enrichment, ^{18}O-tagging, and dephosphorylation are quantitative for all phosphopeptides. Preliminary experiments using synthetic phosphopeptides showed some variability in IMAC affinities, with often higher retention yields for acidic than for basic phosphopeptides (results not shown). In view of this variability, it is possible that phosphorylation sites which are located within basic sequences such as the c-AMP- and c-GMP-dependent protein kinase consensus phosphorylation sites may have been selectively lost in the IMAC enrichment procedure. Similarly, small hydrophilic phosphopeptides could have passed in the flow-through of the RP columns [36]. This argues for the use of a multi-step approach in which the first enrichment step could include an ion-exchange adsorption [10, 11] or an immunoprecipitation step. Alternatively, the use of proteolytic enzymes with specificities other than basic residues could be foreseen. Stable isotope tagging is done by post-cleavage trypsin-catalyzed oxygen exchange at the carboxyl groups of arginine and lysine. We used this strategy because it can be carried out in conditions optimized for oxygen exchange rather than for hydrolysis of peptide bonds [21]. This tagging procedure adds a mass of 4 amu, which is sufficient to distinguish single (phosphorylated) from double (most likely contaminating peptides) isotope envelopes. With different synthetic phosphopeptides, including phosphotyrosine peptides, we noticed complete dephosphorylation by the cocktail of phosphatases. Still, it remains possible that in the large set of cellular phosphopeptides there are components with partial or complete resistance to phosphatase activity. In the latter case, no shifts will be produced at any level and such phosphopeptides will not be sorted in the COFRADIC procedure. In cases where dephosphorylation does not proceed to completion, ^{16}O/^{18}O-doublet envelopes are expected to appear in which the ^{16}O-variant is dominant. The peptide ion with m/z value 670.78 Th shown in Fig. 3 could be such an example as it has a ^{16}O/^{18}O-envelope and therefore scored as a non-phosphorylated peptide. However, the corresponding sequence covered in the DEAD-box protein 20 contains a casein kinase II site predicted by PhosphoBase [37]. This example illustrates the measures taken to avoid assignment of false positives in our studies. As a consequence, Supplementary Table 1 must be considered as a minimal list.

Most identified ex-phosphorylated peptides belong to proteins that are known to be phosphorylated; however, for many of them the exact phosphorylation sites have not yet been mapped. We noticed that the phosphorylation sites of 47 peptides (about 25% of all identified ex-phosphorylated peptides) have been assigned in experiments conducted by other research groups (most notably those described in [10, 11], Supplementary Table 1). This not only indicates that our approach selects for *in vivo* phosphorylated peptides but also points to the fact that our list contains many peptides of which the exact phosphorylation sites can be determined in further experiments.

Our list contains 18 COOH-terminal peptides of 17 different proteins. COOH-termini are often targets for phosphorylation, regulating activities or protein interactions (*e.g.*, [38]). However, because these peptides do not end on a basic residue, they cannot be tagged with ^{18}O-isotopes. Yet they appear as ^{16}O-peptides like the sorted dephosphorylated peptides. However, since they were recovered in the shifted interval, they are likely to be true ex-phosphopeptides. At least in two instances, we could directly prove the presence of phosphorylated residues in these *C*-terminal peptides. For the 60S acidic ribosomal protein P1, we identified serine-104 as the target residue, while for the CTP-synthase, we directly confirmed phosphoserine on sites 571, 573, and 574 (Table 1, entry 4 of the phosphoserine-containing peptides). In other studies, phosphorylation events occurring at the *C*-termini of the 60S ribosomal protein P0 and the tubulin α-1 chain have been identified (Supplementary Table 1). Taken together these findings suggest that the sorted *C*-terminal peptides identified here may not be simply discarded but most likely point to real phosphorylation events occurring in HepG2 cells.

Table 1 sums up those ex-phosphorylated peptides that contain only one possible phosphorylation site and/or of which the phosphorylated counterpart was identified following direct MS/MS analysis of the primary fraction. In total, 38 novel phosphorylation sites were discovered, the majority of them serines. However, five peptides phosphorylated on tyrosine residues were also identified. One interesting example is the EGF receptor. Although phosphorylation on tyrosines 1092, 1110, 1172, and 1197 has been illustrated in the past, here we add a novel site to this phosphorepertoire: phosphorylation on tyrosine-978, which also resides in its cytoplasmic domain and could play a role in docking known target molecules (*e.g.*, the SH3 domain containing CBL proteins) or attract other, yet undiscovered, proteins to the receptor's *C*-terminal domain. While information in the literature on tyrosine phosphorylation of the four other proteins is scarce, it was found that the chicken embryo elongation factor 2 was phosphorylated on tyrosine next to threonine phosphorylation [39]. Here, our findings do indeed hint to phosphorylation on tyrosine-265 of the elongation factor 2.

Most of the results obtained so far are novel. The biological implications of these identified sites will need to be analyzed in more detailed studies. However, we have displayed a powerful tool for detecting and also for measuring phosphorylations of proteins in a detailed manner in highly complex mixtures such as complete cell lysates.

The authors would like to thank Prof. Dr. Jan Tavernier for providing the HepG2 cells. K.G. is a Postdoctoral Fellow and L.M a Research Assistant of the Fund for Scientific Research – Flanders (Belgium) (F.W.O. – Vlaanderen). The project was further supported by a research grant from the Fund for Scientific Research – Flanders (Belgium) (project number G.0008.03), the Inter University Attraction Poles (IUAP, project number P5/05) and the GBOU-research initiative (project number 20204) of the Flanders Institute of Science and Technology (IWT).

5 References

[1] Besant, P. G., Tan, E., Attwood, P. V., *Int. J. Biochem. Cell Biol.* 2003, *35*, 297–309.

[2] Chen, C. C., Bruegger, B. B., Kern, C. W., Lin, Y. C. *et al.*, *Biochemistry* 1977, *16*, 4852–4855.

[3] Levy-Favatier, F., Delpech, M., Kruh, J., *Eur. J. Biochem.* 1987, *166*, 617–621.

[4] Kang, Y. J., Olson, M. O., Jones, C., Busch, H., *Cancer Res.* 1975, *35*, 1470–1475.

[5] Vandekerckhove, J., Bauw, G., Puype, M., Van Damme, J., Van Montagu, M., *Eur. J. Biochem.* 1985, *152*, 9–19.

[6] Pappin, D. J., Højrup, P., Bleasby, A., *Curr. Biol.* 1993, *3*, 327–332.

[7] McLachlin, D. T., Chait, B. T., *Curr. Opin. Chem. Biol.* 2001, *5*, 591–602.

[8] Ficarro, S. B., McCleland, M. L., Stukenberg, P. T., Burke, D. J. *et al.*, *Nat. Biotechnol.* 2002, *20*, 301–305.

[9] Nuhse, T. S., Stensballe, A., Jensen, O. N., Peck, S. C., *Mol. Cell. Proteomics* 2003, *2*, 1234–1243.

[10] Ballif, B. A., Villen, J., Beausoleil, S. A., Schwartz, D., Gygi, S. P., *Mol. Cell. Proteomics* 2004, (Epub ahead of print).

[11] Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E. *et al.*, *Proc. Natl. Acad. Sci. USA* 2004, *101*, 12130–12135.

[12] Shu, H., Chen, S., Bi, Q., Mumby, M., Brekken, D. L., *Mol. Cell. Proteomics* 2004, *3*, 279–286.

[13] Zhou, H., Watts, J. D., Aebersold, R., *Nat. Biotechnol.* 2001, *19*, 375–378.

[14] Goshe, M. B., Conrads, T. P., Panisko, E. A., Angell, N. H. *et al.*, *Anal. Chem.* 2001, *73*, 2578–2586.

[15] Oda, Y., Nagasu, T., Chait, B. T., *Nat. Biotechnol.* 2001, *19*, 379–382.

[16] Amoresano, A., Marino, G., Cirulli, C., Quemeneur, E., *Eur. J. Mass Spectrom.* 2004, *10*, 401–412.

[17] Gevaert, K., Van Damme, J., Goethals, M., Thomas, G. R *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 896–903.

[18] Gevaert, K., Ghesquiere, B., Staes, A., Martens, L. *et al.*, *Proteomics* 2004, *4*, 897–908.

[19] Gevaert, K., Goethals, M., Martens, L., Van Damme, J. *et al.*, *Nat. Biotechnol.* 2003, *21*, 566–569.

[20] Mouri, H., Sakaguchi, K., Sawayama, T., Senoh, T. *et al.*, *Acta Med. Okayama* 2002, *56*, 309–315.

[21] Staes, A., Demol, H., Van Damme, J., Martens, L. *et al.*, *J. Proteome Res.* 2004, *3*, 786–791.

[22] Annan, R. S., Carr, S.A., *J. Protein Chem.* 1997, *16*, 391–402.

[23] Lee, C. H., McComb, M. E., Bromirski, M., Jilkine, A. *et al.*, *Rapid Commun. Mass Spectrom.* 2001, *15*, 191–202.

[24] Tan, Y., Demeter, M. R., Ruan, H., Comb, M. J., *J. Biol. Chem.* 2000, *275*, 25865–25869.

[25] Nigg, E. A., Gallant, P., Krek, W., *Ciba Found. Symp.* 1992, *170*,72–84.

[26] Rossi, R., Villa, A., Negri, C., Scovassi, I. *et al.*, *EMBO J.* 1999, *18*, 5745–5754.

[27] Stokoe, D., Engel, K., Campbell, D. G., Cohen, P., Gaestel, M., *FEBS Lett.* 1992, *313*, 307–313.

[28] Colwill, K., Feng, L. L., Yeakley, J. M., Gish, G. D. *et al.*, *J. Biol. Chem.* 1996, *271*, 24569–24575.

[29] Graff, J. M., Rajan, R. R., Randall, R. R., Nairn, A. C., Blackshear, P. J., *J. Biol. Chem.* 1991, *266*, 14390–14398.

[30] Lees, J. A., Buchkovich, K. J., Marshak, D. R., Anderson, C. W., Harlow, E., *EMBO J.* 1991, *10*, 4279–4290.

[31] Campbell, D. H., Sutherland, R. L., Daly, R. J., *Cancer Res.* 1999, *59*, 5376–5385.

[32] Huang, C., Liu, J., Haudenschild, C. C., Zhan, X., *J. Biol. Chem.* 1998, *273*, 25770–25776.

[33] Marie-Cardine, A., Kirchgessner, H., Eckerskorn, C., Meuer, S. C., Schraven, B., *Eur. J. Immunol.* 1995, *25*, 3290–3297.

[34] Posewitz, M. C., Tempst, P., *Anal. Chem.* 1999, *71*, 2883–2892.

[35] Bonenfant, D., Schmelzle, T., Jacinto, E., Crespo, J. L. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003, *100*, 880–885.

[36] Larsen, M. R., Graham, M. E., Robinson, P. J., Roepstorff, P., *Mol. Cell. Proteomics* 2004, *3*, 456–465.

[37] Kreegipuu, A., Blom, N., Brunak, S., *Nucleic Acids Res.* 1999, *27*, 237–279.

[38] Maudoux, O., Batoko, H., Oecking, C., Gevaert, K. *et al.*, *J. Biol. Chem.* 2000, *275*, 17762–17770.

[39] Kim, Y. W., Kim, C. W., Kang, K. R., Byun, S. M., Kang, Y. S., *Biochem. Biophys. Res. Commun.* 1991, *175*, 400–406.

2.3. Spectrum quality assignment: a priori and a posteriori filtering

2.3.1. Introduction

One of the simplest ways to optimize the efficiency of almost any analysis is to concentrate on the good data produced and to disregard the poor data. In proteomics, concentrating on good data can be achieved by differentiating between high quality and low quality spectra. Interestingly, one of the simplest ways to detect potentially erroneous peptide assignments is to differentiate between identifications from low quality spectra and those from high quality spectra. The only difference between the pre-identification filtering and the post-identification filtering lies in the stringency with which the spectra are classified. A pre-filter will be configured to operate at low stringency; we would rather have 15% of the poor spectra leaking through than suffer the erroneous removal of 1% of the good spectra. Post-identification filtering can be stringent. Here we would not mind being presented with 10% of the really good identifications to reconsider, as long as more than 99% of the poor identifications are also labeled as suspect.
In order to allow the filtering of spectra, an application was built in collaboration with Kristian Flikka and Professor Ingvar Eidhammer of the University of Bergen in Norway. The details of the approach and the corresponding results are outlined in the published paper below.
It is of note that after this paper was accepted in Proteomics, two other papers on spectrum quality filtering were published by Nesvizhskii *et al.* [Nesvizhskii 2005] and Salmi *et al.* [Salmi 2006].

2.3.2. Publication

DOI 10.1002/pmic.200500309 *Proteomics* 2006, *6,* 2086–2094

RESEARCH ARTICLE

Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering

Kristian Flikka[1, 2, 3], Lennart Martens[4, 5]*, Joël Vandekerckhove[4, 5], Kris Gevaert[4, 5] and Ingvar Eidhammer[3]*

[1] Computational Biology Unit, Bergen Center for Computational Science, Bergen, Norway
[2] Proteomics Unit at University of Bergen (PROBE), Bergen, Norway
[3] Department of Informatics, University of Bergen, Bergen, Norway
[4] Department of Medical Protein Research, Flanders Interuniversity Institute for Biotechnology, Department of Biochemistry, Ghent University, Ghent, Belgium

In contemporary peptide-centric or non-gel proteome studies, vast amounts of peptide fragmentation data are generated of which only a small part leads to peptide or protein identification. This motivates the development and use of a filtering algorithm that removes spectra that contribute little to protein identification. Removal of unidentifiable spectra reduced both the amount of computational and human time spent on analyzing spectra as well as the chances of obtaining false identifications. Thorough testing on various proteome datasets from different instruments showed that the best suggested machine-learning classifier is, on average, able to recognize half of the unidentified spectra as bad spectra. Further analyses showed that several unidentified spectra classified as good were derived from peptides carrying unanticipated amino acid modifications or contained sequence tags that allowed peptide identification using homology searches. The implementation of the classifiers is available under the GNU General Public License at http://www.bioinfo.no/software/spectrumquality.

Received: May 9, 2005
Revised: September 28, 2005
Accepted: October 12, 2005

Keywords:
Peptide-centric proteomics / Protein identification / Spectrum quality / Tandem mass spectrometry

Correspondence: Kristian Flikka, Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, Hoeyteknologisenteret, Thormoehlensgate 55, N-5008 Bergen, Norway
E-mail: flikka@ii.uib.no
Fax: +47-55584295

Abbreviations: AUC, area under ROC curve; **AODE**, aggregating one-dependence estimators; **COFRADIC**, combined fractional diagonal chromatography; **LBR**, Lazy Bayesian Rules; **NB**, naïve Bayes; **ROC**, receiver operator characteristics; **SP-TAN**, super parent tree augmented naïve Bayes; **TAN**, tree augmented naïve Bayes; **TN**, true-negative

1 Introduction

When analyzing a complex peptide mixture by automated one-or multidimensional LC MS/MS, a vast number of peptide fragmentation spectra are typically obtained. These are linked to peptides/proteins using popular search algorithms such as MASCOT [1] and SEQUEST [2]. Nevertheless, a significant number of MS/MS spectra remain unidentified, due to different reasons; there may be too little fragment ion

* These authors contributed equally to this work.

information in the spectrum, the fragmented precursor may not be a peptide, the peptide may be modified in a way that is unaccounted for by the search algorithm, or the peptide may not be present in the searched database. Successful identification further depends on good preprocessing of mass spectrometric data [3].

With the advent of several techniques for sensitive proteome analysis [4, 5], the importance of automatic methods for spectrum preprocessing has increased. Manual validation of all resulting spectra in such studies is not feasible and tools enabling an automatic quality control are therefore important as an integrated part of a modern high-throughput proteomics laboratory [3, 6].

Recently, a study was published [7] describing methods for automatic spectrum quality assessment. In this paper, two different classification schemes were evaluated: one using handcrafted attributes and another using support vector machines (SVM) based on observed m/z values. A collection of spectra from a mixture of five known proteins digested with four different proteases [8] was used to test the classification algorithms. The spectra were initially identified using SEQUEST [2], and identified spectra were labeled "good", whereas all other spectra were labeled "bad". When testing the SVM-based classifier, 90% of the identified spectra were labeled good and 75% of the unidentified ones were labeled bad; thus, 75% of the unidentified spectra could be removed, while removing 10% of the identified ones.

Another study was recently published, presenting the program SPEQUAL [9], in which spectra from 23 peptides and 12 proteins were analyzed with respect to their quality, based on three filter components. In particular, it was evaluated whether a spectrum's charge was correctly assigned, a score based on the sum of all peak intensities was applied (TIC) and an S/N score was finally calculated. To test this application, 10 000 spectra were manually given a quality label, "low", "intermediate", or "high". The labels were based on the same three criteria as the algorithm itself was composed of. The spectra were then run through the program, and the output score was compared against the manually assigned labels. A good correspondence between the program score and the assigned labels was observed. The results from the program were however not compared to the results from database searches; thus, the number of identifications that are lost using the SPEQUAL procedure, was not explicitly stated.

In our study, we adopted the notation of spectra, identified by MASCOT, as good, and unidentified as bad from [7]. Furthermore, we employ a promising classifier based on aggregation of one-dependence Bayesian classifiers [10] testing this on data from different mass spectrometers using different sets of samples that were all generated by peptide-centric combined fractional diagonal chromatography (COFRADIC) [11–13]. Various other Bayesian and decision tree classifiers were tested and compared to the aggregating one-dependence estimators (AODE) method. We briefly address the issue of spectrum clustering, where a dataset is reduced to containing only nonequivalent spectra, and demonstrate the necessity of performing clustering on the data in machine-learning studies of proteomic data. Removing redundant spectra decreases the potential risk of overfitting the models to highly abundant peptides and thus enables a reliable cross-validation of the results. Results from the classification test using ESI IT fragmentation data showed that 83% of the bad spectra could be removed, while removing only 10% of the good ones. The corresponding removal rate for clustered data was 79%. The fact that the removal rate is lower for the clustered data suggests that the original redundancy in the data gives overoptimistic results; hence, a clustering is crucial to get reliable test results.

By using the classifiers suggested in this article in conjunction with a proteomics mass-spectrometry pipeline for identification of proteins, we were able to reduce the number of spectra needed for the analysis, detect false identifications, and suggest a number of high-quality unidentified spectra that were subsequently de novo sequenced.

2 Materials and methods

The data analysis was performed on peak lists generated by the mass spectrometer vendor's proprietary software using the raw peptide fragmentation spectra. In order to devise a powerful classifier, it was found important to use datasets for training and testing that were realistic in terms of both size and complexity.

Similar to Bern *et al.* [7], we denoted the spectra identified by the MASCOT algorithm [1] as good, all others as bad.

2.1 Experimental data

We used four different proteome datasets that were obtained following LC-MS/MS or MALDI-TOF-TOF analysis of peptides isolated by two different strategies. Three different mass spectrometers were used in these studies (see below). Identification was done using MASCOT and all identifications were manually verified. Spectra that scored equal to or above the MASCOT identity threshold at the 95% confidence level were accepted. All spectra and identifications are available at the PRIDE [14] server (http://www.ebi.ac.uk/pride).

2.1.1 Q-TOF N-terminal dataset

These MS/MS spectra are from the extreme amino terminal peptides of proteins which were isolated as described in [13]. A total of 10 054 spectra were available for this study. These spectra were incrementally searched against the IPI database containing only human proteins [15] and IPI-derived, N-terminally truncated databases as described in [13] (PRIDE accession number 1643).

2.1.2 Q-TOF metOx dataset

This dataset consists of MS/MS spectra from tryptic methionyl peptides as described in [13] and were searched against the human IPI database. Here, 3565 spectra were studied (PRIDE accession number 1644).

2.1.3 IT dataset

The data (7477 spectra) were obtained from an N-terminal COFRADIC analysis of the human SH-SY5Y cell line. The IT instrument used was a Bruker Esquire HCT (Bruker Daltonics GmbH, Bremen, Germany). Sample preparation and data processing were performed as described in [12]. These spectra were incrementally searched against the human subset of the Swiss-Prot (ftp://us.expasy.org/databases/swiss-prot) and NCBI nr (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/protein) databases and hereof derived, N-terminally truncated databases as described in [13] (PRIDE accession number 1645).

2.1.4 MALDI-TOF-TOF dataset

PSD spectra were obtained using an Ultraflex MALDI-TOF/TOF mass spectrometer (Bruker) operating in an automated analysis mode. Peptides used for this analysis were methionyl peptides isolated as their sulfoxide derivatives [11] from a tryptic proteome digest of human Jurkat T-lymphocytes and multipotent adult progenitor cells (MAPC). For this study, a total of 8338 spectra were used, of which 3771 were made available (corresponding to the Jurkat dataset). The PRLDE accession number for the latter is 1646.

2.2 Extracting attributes

A variety of attributes may be considered relevant when evaluating a spectrum's quality, see [16]. Given a spectrum, it should be possible to extract the value of any defined attribute and to represent it as a numerical value. Assume that if we have k attributes, a spectrum is represented by a feature vector $x = \langle x_i \ldots x_k \rangle$, where x_i is the value of the ith attribute. Furthermore, we seek to classify a spectrum into one of two classes, denoted y, where $y \in \{good, bad\}$.

A selection of attributes was used to capture the variety of spectral features in the data material. Some of them are classic; some are new. The attributes may be divided into two categories; automatically specified and manually specified. The automatic specification of attributes considers all possible between-peak mass differences (deltas) and all possible m/z values. All the observed m/z values and deltas are scored using the chi-square analysis for contingency tables, and only the highest scoring ones are used as attributes. This is done by counting the number of occurrences for each m/z value, for both the good and the bad spectra. If an m/z value is significantly over-represented in either the good or bad spectra, the chi-square test will reveal this.

The manually specified attributes are listed in Table 1. Some of these are commonly known, and can also be found in the literature [7, 16], such as number of peaks, total peak intensity, and total intensity of complement fragment ions. Most of them are, however, other attributes, suggested by experienced mass spectrometer operators. We defined the relative intensity of each peak as the peak's intensity divided by the intensity of the highest peak. When specified, we only considered peaks with a relative intensity above a certain threshold; we used 0.1. This threshold was empirically found to produce the overall best results. Using the rank-based intensities suggested in [7] did not improve the results.

Combining the manually and automatically specified attributes resulted in maximum 78 features, fewer if the automatically specified features were not found significantly different between the identified and unidentified spectra. To avoid bias in the group of attributes, only the training spectra were used as basis for feature selection.

2.3 Spectrum clustering

In several articles [17, 18] it has been shown that large collections of tandem mass spectra often carry significant redundancy. Different spectra that most likely represent the same peptide will introduce overoptimistic results when testing a classifier, and may also bias the classifier itself toward the most frequently occurring spectra. For these reasons we have trained and tested all methods on both full and reduced datasets, where the reduced ones are the results of a clustering based on the principles described in [18].

A possible drawback of collapsing several spectra into one may be the loss of information about the relative importance of certain types of spectra, in particular for some contaminants. Typically, these consist of non-biopolymers (such as PEG) that have quite similar fragmentation spectra though different precursor masses. Since the precursor mass needs to match up for spectra to be clustered, these contaminant spectra will not be collapsed and therefore the information regarding abundance will not be lost.

2.4 Lining up an ensemble of machine-learning methods

The recently developed Bayesian classifier AODE described in [10] was put to the test against an assembly of methods, some of which were also used in the article by Webb *et al.* [10].

Generally, probability-based classifiers will estimate $P(y|x)$ for each value of y, *i.e.*, the probabilities of a spectrum belonging to class good or bad given the spectrum's feature vector x. Selecting the class y that maximizes an estimate $\hat{P}(y|x)$ will then be the classification from the method. Using Bayes theorem, the problem can be reduced to maximizing $\hat{P}(x|y)\hat{P}(y)$. This is not straightforward, as a direct estimate for $P(x|y)$ from training data requires a significant number of observations for each possible x. For example, having

Table 1. Overview of manually specified attribute variables

Feature id	Description
num_peaks	Number of peaks in spectrum
num_sign_peaks	Number of peaks with relative intensity >0.1
rel_num_sign_peaks	Number of significant peaks divided by precursor mass
avg_delta_mass	The average delta mass in spectrum
std_dev_delta_mass	SD of delta mass values
Precursor_charge	Charge of precursor ion
precursor_mass	Mass of uncharged precursor
precursor_mz	*m/z* Value of precursor in parent spectrum
rel_intºf_prec_in_msms	Relative intensity of precursor in fragment spectrum
delta_two_highest	Intensity difference between top two peaks
avg_peak_density	Number of peaks/(max_mz – min_mz)
num_dominant_peaks	Number of peaks accounting for >5% of total intensity
avg_rel_peak_intens	The average of relative peak intensities
std_dev_rel_peak_intens	The SD of relative peak intensities
raw_total_intensity_threshed	Total raw intensities for significant peaks
rel_total_intensity_threshed	Total relative intensities for significant peaks
rel_int_by_compl_to_prec	Total relative intensity of complement pairs

Significant peaks have relative intensity above 0.1.

50 attributes, each allowing ten different values, may produce up to 10^{50} different *x*-vectors. To circumvent this, the naïve Bayes (NB) [19] algorithm makes the assumption that the attributes are independent of each other, only given the class. Hereby follows that: $P(x|y) = \prod_{i=1}^{k} P(x_i|y)$ which is what the classifier NB uses, and thus classifies an instance by choosing the class that maximizes the probability estimates: $\underset{y}{argmax}\left(\widehat{P}(y)\prod_{i=1}^{k}\widehat{P}(x_i|y)\right)$.

The NB algorithm is very fast, and delivers optimal classification when its constraint is satisfied. It performs quite well, even when the underlying independence assumption is clearly unrealistic. Significant work has been done to weaken the attribute independence assumption; examples of this are: Lazy Bayesian Rules (LBR) [20], Tree Augmented NB (TAN) [21], Super parent TAN (SP-TAN) [22], and the recent algorithm, AODE [10]. In the latter article it is suggested that the AODE algorithm has performance comparable to that of LBR and SP-TAN, but at a significantly lower computational cost.

The idea of AODE is to aggregate the predictions made by a collection of one-dependence classifiers. These classifiers require each attribute to be depending on exactly one other attribute in addition to the class. The difference in computational cost between the NB and the AODE algorithm can be illustrated as adding an extra dimension to the table of probability estimates. For NB, this table is indexed by the target attribute value and the class value. In AODE, there is an extra dimension indexed by the value of the attribute that the target is conditioned on.

To this end, we tested the Bayesian methods NB, AODE, SP-TAN, LBR, TAN, and ODE, where SP-TAN and LBR both have been shown to generate relatively small errors in their predictions. Their drawback is the slow training/testing execution, in particular for the LBR method. In this study, we had to discard the LBR classifier, because of its prohibitively long testing time. The ODE is equal to AODE, apart from the aggregation of one-dependence estimators. Where AODE aggregates one-dependence estimators, ODE only uses a single one.

Decision trees, often represented by the C4.5 algorithm [23], are used for a large variety of machine-learning tasks. One of the advantages with decision trees is the fact that the resulting trees can be examined to discover the features that are most important. The alternating decision tree algorithm [24] uses a data structure called ADTrees to represent decision tree classifiers. Both the ADTree and the C4.5 algorithm were tested on all datasets and compared to the other methods.

2.5 Implementation

The classifiers have been thoroughly tested using five-fold cross-validation. In addition, all the datasets have been reduced to containing only spectra with little internal similarity using a procedure similar to the one described in [18]. The implementation of the classification systems was done in Java (http://java.sun.com) using the open source package WEKA [25] (http://www.cs.waikato.ac.nz/ml/weka/). We used the implementation by Webb *et al.* [10] for the SP-TAN and ODE classifiers. For all methods, WEKA's default parameter values were used. The reimplementation of C4.5 is

denoted J48, and we applied boosting to enhance the performance of the J48 algorithm. WEKA's implementation of AdaBoost M1 [26] was used for boosting.

For the Bayesian classifiers, the attribute values were discretized by applying the MDL method [27]. In the performance evaluations the spectra classified as good spectra are denoted as the positives, the others as negatives.

3 Results

3.1 Spectrum quality classifier

An overview of the results can be found in Table 2 and Fig. 1. As seen in Fig. 1, the classifier performance drops when applied to the reduced versions of the datasets. This may be caused by the fact that the presence of similar spectra in the training- and test-sets corrupts the testing scheme and emphasizes the importance of a stringent strategy when testing machine-learning methods. The difference in performance between the full and reduced datasets is small on the MALDI-TOF-TOF instrument and on the IT instrument. The nature of the MALDI-TOF-TOF experiments leads to few fragmentations of the same peptide, which again reduce the number of similar spectra. For the IT instrument the explanation is most likely an improved dynamic mass-exclusion strategy (as compared to the Q-TOF), which aims at reducing the chance of a peptide being fragmented repeatedly in a short time frame. From this point on we address performance figures only on the reduced datasets, unless stated otherwise.

The error rates in these experiments are heavily influenced by the typical imbalanced nature of the data. In fact, the number of unidentified (bad) spectra was typically six times higher than the number of identified (good) spectra. Thus, a simple classifier assigning all spectra as bad gives an error rate of $1/7 = 0.143$, which is relatively good. Using receiver operator characteristics (ROC) curves and the area under ROC curve (AUC) in this experiment is a more suitable measure of performance, because this also gives an impression of how well a classifier can be adjusted to a certain target, if we for example only accept classifiers with a true-positive rate of 0.98 or better.

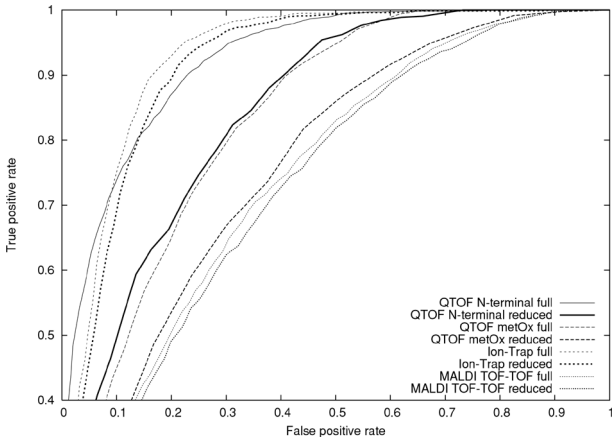


Figure 1. ROC curves for all datasets, clustered and unclustered using the AODE classifier. Difference in performance is illustrated by the fact that all datasets have better ROC curves for the unclustered versions of the data.

Motivated by the high dimensionality of the data (many attributes), various machine-learning methods often select some attributes as more important than others. A useful method for evaluating the importance of various attributes is to calculate the information gain [28]. In Table 3, the ten highest scoring attributes for each dataset are listed. The manually specified attributes dominate the lists, in particular "tot_rel_int_by_compl_to_prec", the total relative intensity of pairs of peaks whose masses together correspond to the precursor mass. Automatically specified attributes in the list include among others "countDelta113.1" (leucine/iso-

Table 2. Overall results for all methods and datasets. The table shows the AUC and the true-negative (TN) rate for each dataset on each method. AUC is the area under the ROC curve. TN rate is the rate of TNs, *i.e.*, unidentified spectra classified as bad, given a true-positive rate (rate of identified spectra labeled good) of 0.90. AODE simple is a version of the AODE classifier using only 10 attributes

Method	Q-TOF N-terminal (3265 spectra)		Q-TOF metOx (2006 spectra)		IT (5948 spectra)		MALDI–TOF-TOF (7013 spectra)	
	AUC	TN rate	AUC	TN rate	AUC	TN rate	AUC	TN rate
AODE	0.84	0.60	0.76	0.45	**0.91**	**0.79**	**0.73**	**0.38**
ODE	0.79	0.52	0.75	0.41	0.88	0.75	0.71	0.35
NB	0.79	0.52	0.75	0.40	0.88	0.74	0.71	0.35
TAN	0.83	**0.61**	**0.77**	**0.48**	0.90	0.78	**0.73**	0.36
SP-TAN	0.63	0.28	0.73	0.39	0.68	0.22	0.67	0.32
ADTree	0.82	0.54	0.75	0.41	0.89	0.76	0.70	0.32
J48	**0.85**	0.55	0.70	0.33	0.89	0.75	0.69	0.30
AODE simple	0.82	0.57	0.74	0.38	0.90	0.76	0.71	0.31

Table 3. Importance of each attribute ranked using the information gain [28] measure. Each dataset has one column in the table, and the top ten ranked attributes are shown sorted by importance. Datasets were reduced to containing only one instance of each spectrum

Q-TOF N-terminal	Q-TOF metOx	IT	MALDI-TOF-TOF
tot_rel_int_by_compl_to_prec	tot_rel_int_by_compl_to_prec	tot_rel_int_by_compl_to_prec	raw_total_intensity_threshed
precursor_mass	precursor_mass	Num_peaks	std_dev_rel_peak_intens
raw_total_intensity_threshed	precursor_mz	avg_delta_mass	avg_delta_mass
num_peaks	countDelta1.0	std_dev_delta_mass	num_peaks
countDelta113.1	num_peaks	raw_total_intensity_threshed	countDelta113.1
precursor_charge	raw_total_intensity_threshed	num_peaks_to_account_forX_intensity	avg_rel_peak_intens
num_sign_peaks	rawMass133.1	avg_rel_peak_intens	precursor_mz
avg_delta_mass	num_sign_peaks	num_sign_peaks	precursor_mass
std_dev_delta_mass	rawMass89.1	CountDelta113.0	countDelta2.0
countDelta112.1	rawMass177.1	CountDelta170.1	countDelta112.8

leucine), "rawMass89.1", "rawMass133.1", and "rawMass177.1" (fragment peaks of PEG that are most often seen in the unidentified spectra). Attributes "countDelta112.1" and "countDelta112.8" most likely also originate from leucine or isoleucine. The mass difference of 112.1 can be expected to occur together with the correct difference at 113.1 because the fragments are often represented with more than one isotope, whereas the difference of 112.8 may stem from inaccurate measurements. "countDelta170.1" may correspond to a mass difference spanning two amino acids; matching duplets include leucine/isoleucine plus glycine and alanine plus valine.

The data from the IT instrument appear to give the most beneficial results for most methods. By using the AODE classifier, 62% of the unidentified IT MS/MS spectra can be removed without removing more than 2% of the identified spectra (Fig. 1). Among the tested methods, the AODE classifier gives the highest AUC (0.91), whereas the SP-TAN method only has an AUC of 0.68.

It appears that the effectiveness of the classifiers varies significantly between experiments as some datasets are more difficult to classify for all tested methods. Comparing the results for the two Q-TOF datasets reveals that spectra derived from N-terminal peptides are possible to classify quite well: 45% of the unidentified spectra can be removed while removing only 2% of the identified ones, using the AODE classifier. Q-TOF MS/MS spectra from methionyl peptides are however less separable: only 25% of the unidentified spectra can be removed when removing 2% of the identified ones. The MALDI-TOF-TOF data results in very few removable, unidentifiable spectra: 20% of the unidentified spectra are recognized as bad, when removing only 2% of the identified ones.

We believe that these spectral differences might be due to the fact that some experimental procedures and mass spectrometers generally generate fewer bad and unidentifiable spectra as compared to other setups.

Difficulties in classifying MALDI-TOF-TOF data are most likely caused by the overall high quality of the spectra generated by this instrument. Indeed, when operating an ESI-based mass spectrometer in an automated LC-MS/MS mode, the instrument quite randomly and continuously picks ions for subsequent fragmentation analysis, see [29]. As many contaminants (polymers, detergents) tend to ionize quite easily, they are frequently picked up and analyzed further by MS/MS, thereby inevitably resulting in easily recognizable bad MS/MS spectra. MALDI-MS does not suffer from this drawback as samples are archived and intelligent peak picking algorithms are used that only mark ions that are most likely derived from peptides and/or are intense enough such that good-quality fragmentation spectra might be generated. In particular, the instrument software is set to only consider ions with an S/N of 60, rank them according to the quality of their isotopic envelope, and submit only the best of these to MS/MS analysis. Hence the general and overall spectrum quality difference between MALDI and ESI peptide fragmentation spectra.

More subtle differences as those indicated above for the Q-TOF MS/MS spectra of amino terminal and methionyl peptides might originate from the number of peptides that were available for analysis. Clearly, isolated N-terminal peptides offer the highest possible reduction in analytes to be analyzed in peptide-centric proteome studies as every protein is finally represented by only one peptide: its N-terminal one. On the other hand, on average about one-fifth of all peptides in a tryptic proteome digest contain methionine. This implies that the filling of MS analysis time with real peptides is less when analyzing N-terminal peptides than with methionyl peptides, whereby in the former case mass spectrometers are given an increased opportunity for fragmenting "non-peptides" and thus start filling data space with peak lists from bad spectra eventually making the distinction between good and bad spectra easier (see also [13]).

A version of the AODE classifier that only uses ten features is also included for comparison. The ten features are selected by using the information gain measure. This classifier is a weaker performer than the full AODE classifier, although the difference is relatively small. It may, however, be beneficial to include more features to ensure that the classifier is more robust to possible changes in the characteristics of the datasets.

By examining background data for the ROC curves we find the connection between the true-positive rate (fraction of identified spectra, classified as good) and the false-positive rate (fraction of unidentified spectra classified as good). From Figs. 1, 2 we can see that by introducing a penalization of the false negatives (good spectra classified as bad), we can shape the classifier to suit different needs. Shaping Bayesian classifiers is obtained by changing the prior class probabilities. If originally there is a 9:1 relationship between the bad and the good spectra, the prior class probabilities will be 0.9 and 0.1, respectively. Changing the prior probability of the good ones to 0.2 will result in fewer misclassified good spectra.

When retaining 90% of the identified spectra, it appears that between 38 and 79% of the bad spectra can be removed, depending on the dataset (see Table 2). Another scenario is to fix the true-positive rate at ~0.98, *i.e.*, only accept to falsely remove 2% of the good spectra. Then, the fraction of bad spectra that can be removed drops to between 20% and 62% (not shown in table). It is worth having in mind that when using a *p*-value cut-off on 0.05 using for example MASCOT database search, we may expect up to 5% false positives. Thus, we should not expect the classifier to be able to give a true-positive rate of 1.0, as some of the identified (good) spectra may be false identifications.

In order to choose the overall best classifier, we compared the AUC values for all methods on all datasets, see Table 2. The overall winner is the AODE classifier, as it is victorious on two of the datasets, and close to best on the other two datasets as well. The poor performance of the SP-TAN method is somewhat surprising, but it appears that the high number of attributes may be causing some of the problems. When using few attributes (results not shown), SP-TAN was more on par with the other methods.

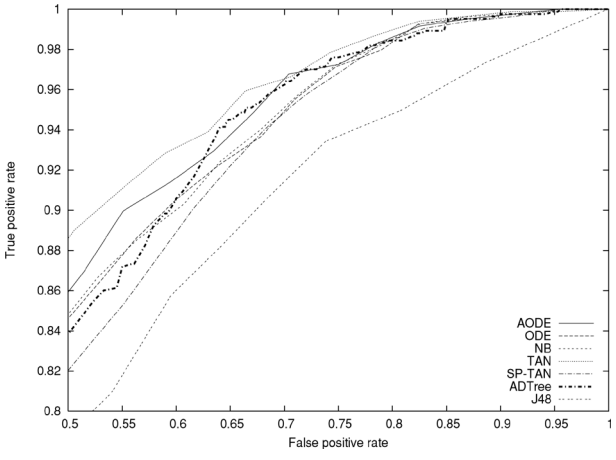


Figure 2. ROC curves for a selected dataset (QTOF metOx), to illustrate the strength of different methods in different areas of the ROC domain. Between TP rate 0.86 and TP rate 0.94 the ADTree method goes from being fifth best to second best.

On the Q-TOF metOx data, the performances of the methods are somewhat similar. There is, however, a difference that can be seen in the ROC curves (Fig. 2). In the low true-positive rate region the ADTree method is quite mediocre, but when true-positive rate rises between 0.86 and 0.94, it becomes almost on par with the best method. The ROC curve of ADTree is, in other words, steeper in this region. This illustrates the fact that before choosing a particular method, the desired strictness of the classifier should be established.

Training times and testing times of the classifiers are different for the various methods. An illustrative example of execution times can be found in Table 4, using the Q-TOF N-terminal dataset. For the Bayesian classifiers, the time used for discretization is added to both training and testing times. The AODE classifier was, for example, about five to ten times faster than ADTree on training time, but ADTree classifies instances extremely fast, so AODE could not compete in the testing phase. The SP-TAN and the J48 classifiers are by far the slowest on the training stage, but classify rapidly. All other methods have execution times that should be recognized as acceptable.

Table 4. Running time examples for all the methods on the Q-TOF N-terminal dataset. Times are expressed in seconds. Numbers should at most be used as a guide to the relative relationship between the methods, as the actual times depend on the implementation and runtime environment

Classifier	Training time, s	Testing time, s	Total time, s
AODE	3	7	10
ODE	2.8	5	7.5
NB	2.5	2.5	5
TAN	12	11	23
SP-TAN	354	4	358
ADTree	23	1	24
J48	85	1	86

3.2 Applications of the classifier for proteomics datasets

An important part of this work was to evaluate the potential applications of the spectrum classifier in a high-throughput proteomics workflow. Since the classifier described here was designed to be versatile, we exploited its adaptability in applying it to proteomics datasets both before and after the identification process.

When the classifier is configured to allow very few false-negative spectra (identified spectra labeled bad), it can be used as a spectrum prefilter, effectively removing "junk" spectra (*e.g.*, spectra from contaminants) prior to identification. This reduces the chance of false identifications and simultaneously boosts the identification efficiency. When we applied this strategy to the clustered N-terminal COFRADIC proteome of human blood platelets by allowing only 1.1% false-negative spectra, approximately 30% of all spectra could safely be removed prior to identification. Seven spectra were labeled bad by the classifier, even though they had a MASCOT score above the identity threshold at 95% confidence. We manually verified these seven identifications and found four of them to be false identifications. The net result of using this stringent classifier prior to identification would therefore have been the loss of three good peptide identifications and the omission of four false-positive identifications while the identification efficiency would have been boosted by a factor of 1.43 through the retention of only 70% of the spectra for effective database searching.

If the classifier is configured to be more sensitive to bad spectra (by *e.g.*, allowing for 10% false-negative spectra), it can be used as a post-identification quality control, highlighting those spectra that, although successfully identified by MASCOT at the 95% confidence interval, show signs of potential "badness". Since this subset will be of a more readily manageable size, it can be manually validated, further reducing the final false-positive count.

The classifier can also be applied post-identification to reduce the number of unidentified spectra. These can occur because of the absence of the corresponding peptide sequence from the search database or because of unexpected modifications on the peptide, resulting in aberrant fragmentation patterns as compared to *in silico* predicted patterns. By applying the classifier to the unidentified spectra, we can find those spectra that are classified as good. These can then be submitted to a search against a larger, unrestricted database (*e.g.*, without species restrictions). By applying this strategy to the ion-trap spectra of the N-terminal COFRADIC proteome of neuronal SH-SY5Y cells, we could pick up the known cell culture contaminant BSA, albeit by only two spectra. Human herpes virus and corona virus surface proteins were also picked up.

Further compensation for the absence of matching sequences can be obtained by using automated de novo sequence analysis tools such as Lutefisk [30]. We applied LutefiskXP to 467 spectra from the Q-TOF N-terminal dataset that were unidentified yet marked good by the classifier, corresponding to 18% of the unidentified spectra. The resulting sequence tags were submitted to BLAST and searched against UniProt (www.uniprot.org). Lutefisk suggested 210 high scoring ($Pr(c) > 0.9$) sequences and from these, 24 different sequences gave perfect BLAST matches. These matches identified 20 different human proteins, many of which were abundant in blood platelets as assessed by our previous data [13]. However, several of these proteins were not yet identified in our platelet proteome study and can hence be added to the list of platelet proteins. Examples include the pleckstrin homology (PH) domain containing proteins ARHGEF18 (Q6DD92) and Pleckstrin (Q6FGM8) and the hypothetical proteins FLJ45525 (Q6ZSH5) and Talin-1 (Q9UPX3).

The classifier also serves as a means to extract unexpected modifications from the unidentified spectra. These modifications can stem either from *in vivo* processes or from *in vitro* artifacts. We analyzed the unidentified, yet good spectra from the N-terminal COFRADIC proteome of human blood platelets for mass differences between fragment peaks that corresponded with the modifications listed in UniMod [31], and ranked them by number of occurrences. This crude list was subsequently analyzed by hand to pick out those modifications that seemed plausible based on the chemistry employed, or for their biological significance. Cyclization of N-terminal *S*-carbamoylmethylcysteine (pyro-cmc) [32] and N-terminal carbamylation were selected as potential artifactual modifications and N-terminal formylation along with mono-methylation of lysines were considered as potential *in vivo* modifications. The pyro-cmc modification resulted in 36 additional identified spectra and N-terminal carbamylation was present in 15 spectra, comprising 8 unique peptide sequences. Formylation added 13 identified spectra containing 6 unique sequences, and methylated lysines were picked up in 10 spectra that collapsed into 6 unique peptide sequences. The classifier proved very useful in this approach as it allows spectrum preselection, so only potentially meaningful fragmentation spectra are searched for the mass signatures of known modifications.

Additionally, because of the high prevalence of the pyro-cmc modification, this has since been added as a variable modification in searches for other datasets. In all of these projects, more than 5% of the identified peptides carry this modification, all of which would otherwise have resulted in false negatives.

4 Discussion

Building a quality-based classifier for tandem mass spectrometric data has to take several aspects into consideration. We have created a versatile classifier that can be adapted to different demands. Rebuilding of the model based on new datasets is done within minutes. The thresholds deciding when a spectrum is deemed bad can be altered to suit the individual needs.

Results show that up to 62% of unidentified spectra can be removed before the identification step, without removing more than 2% of the spectra that would have been identified. This removal of bad spectra thus helps to boost the identification efficiency as well as a reduction in false-positive identifications. Additional quality control of the obtained identifications can be performed by selecting the borderline spectra from the identified set and manually validating these. Furthermore, selecting the potentially identifiable spectra from the large group of unidentified spectra after identification yields a reduced dataset

that can be explored by alternative methods of identification such as de novo sequence analysis and BLAST queries. These potential false-negative spectra can also be mined for unexpected modifications or biological contaminants resulting from sample handling or cell culture artifacts.

The fact that our classifier can be fine-tuned to suit specific needs in stringency allows it to be easily adapted to different proteomics techniques, each yielding different levels of protein coverage by the identified peptides. Higher redundancy approaches (such as multidimensional protein identification technology) can allow higher stringency as the inadvertent loss of a few good spectra would have much less impact on the total amount of proteins identified. Methods with lower redundancy (for example N-terminal COFRADIC) can choose to loose less good spectra and would thus tolerate more bad spectra in the search set.

Direct comparison against the study described in [7] was not possible because the data and software used in that study was not available. The program SPEQUAL [9] is available and has been tested on our datasets. There is however a different focus in their study; an example is that spectra without correct charge state assigned are classified as low-quality spectra. For our IT data, 27% of the identified spectra have no charge state assigned, and will thus be deemed low-quality by SPEQUAL. This makes the SPEQUAL procedure somewhat unsuited for the scope of this study.

The suggested versatile spectrum classifier can be an important tool for high-throughput proteomics analyses with applications in many key stages of the identification process.

The authors would like to thank Jozef Van Damme for his critical remarks and expert input during the initial construction phase of the algorithm, and Kari E. Fladmark for initiating the collaboration. K.F. was supported by the FUGE program in The Research Council of Norway (western regional funds). K.G. is a Postdoctoral Fellow and L.M. a Research Assistant of the Fund for Scientific Research – Flanders (Belgium) (F.W.O. – Vlaanderen). The project was further supported by a research grant from the Fund for Scientific Research – Flanders (Belgium) (project number G.0008.03) and the GBOU-research initiative (project number 20204) of the Flanders Institute of Science and Technology (IWT) to the laboratory in Ghent.

5 References

[1] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *18*, 3551–3667.

[2] Eng, J. K., McCormack, A. L., Yates, J. R. III, *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[3] Sadygov, R. G., Cociorva, D., Yates, J. R. III, *Nat. Methods* 2004, *1*, 195–202.

[4] Zhang, H., Yan, W., Aebersold, R., *Curr. Opin. Chem. Biol.* 2004, *8*, 66–75.

[5] Gevaert, K., Vandekerckhove, J., *DDT: Targets* 2004, *3*, S16–S22.

[6] Carr, S., Aebersold, R., Baldwin, M., Burlingame, A. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[7] Bern, M., Goldberg, D., McDonald, W. H., Yates, J. R. III, *Bioinformatics*, 2004, *20*, i49–i54.

[8] MacCoss, M. J., Wu, C. C., Yates, J. R. III, *Anal. Chem.* 2002, *74*, 5593–5599.

[9] Purvine, S., Kolker, N., Kolker, E., *OMICS* 2004, *8*, 255–265.

[10] Webb, G. I., Boughton, J. R., Wang, Z., *Machine Learning* 2005, *58*, 5–24.

[11] Gevaert, K., Van Damme, J., Goethals, M., Thomas, G. R. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 896–903.

[12] Gevaert, K., Goethals, M., Martens, L., Van Damme, J. *et al.*, *Nat. Biotechnol.* 2003, *21*, 566–569.

[13] Martens, L., Van Damme, P., Van Damme, J., Staes, A. *et al.*, *Proteomics* 2005, *5*, 3193–3204.

[14] Martens, L., Hermjakob, H., Jones, P., Adamski, M. *et al.*, *Proteomics* 2005, *5*, 3537–3545.

[15] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y. *et al.*, *Proteomics* 2004, *4*, 1985–1988.

[16] Tabb, D. L., Eng, J. K., Yates, J. R. III, in: James, P. (Ed.), *Proteome Research: Mass Spectrometry*, Springer, Berlin 2001, pp. 125–142.

[17] Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., Yates, J. R. III, *Anal. Chem.* 2003, *75*, 2470–2477.

[18] Beer, I., Barnea, E., Ziv, T., Admon, A., *Proteomics* 2004, *4*, 950–960.

[19] Langley, P., Iba, W., Thompson, K., in: *Proceedings of the 10th National Conference on AI,* San Jose, CA 1992, pp. 223–228.

[20] Zheng, Z., Webb, G. I., *Machine Learning* 2000, *41*, 53–84.

[21] Friedman, N., Geiger, D., Goldszmidt, M., *Machine Learning* 1997, *29*, 131–163.

[22] Keogh, E. J., Pazzani, M. J., *Proceedings of the International Workshop on AI and Statistics* 1999, pp. 225–230.

[23] Quinian, J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA 1993.

[24] Freund, Y., Mason, L., in: Bratko, I., Dzeroski, S. (Eds.), *Proceedings of the 16th International Conference on Machine Learning*, Morgan Kaufmann 1999, 124–133.

[25] Holmes, G., Donkin, A., Witten, I. H., in: *Proceedings Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia 1994, pp. 357–361.

[26] Freund, Y., Schapire, E., in: Saitta, L. (Ed.), *Proceedings of the 13th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, 1996, pp. 148–156.

[27] Fayyad, U. M., Irani, K. B., in: *Proceedings of the 13th International Joint Conference on AI*, Morgan Kaufmann 1993, pp. 1022–1027.

[28] Witten, I. H., Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, San Francisco, CA, 1999, pp. 91–94.

[29] Liu, H., Sadygov, R. G., Yates, J. R. III, *Anal. Chem.* 2004, *76*, 4193–4201.

[30] Taylor, J. A., Johnson, R. S., *Anal. Chem.* 2001, *1*, 2594–2604.

[31] Creasy, D. M., Cottrell, J. S., *Proteomics* 2004, *4*, 1534–1536.

[32] Geoghegan, K. F., Hoth, L. R., Tan, D. H., Borzilleri, K. A. *et al.*, *J. Proteome Res.* 2002, *1*, 181–187.

2.4. PRIDE: sharing data in a scientific community

"Information, no matter how expensive to create, can be replicated and shared at little or no cost."

- Thomas Jefferson

2.4.1. Introduction

One of the basic premises of science since the Enlightenment is that it should be a collaborative effort, building on open communication of published results. After all it was Newton himself who asserted[59]: *"If I have seen further it is by standing on the shoulders of Giants"*.

The PRoteomics IDEntifications (PRIDE) system was designed to alleviate the needs expressed in section 1.3.4 and the unofficial motto of PRIDE is therefore appropriately '*making publicly available data publicly accessible*'.

The need for the PRIDE system, its adherence to standards, data structure, interface and (potential) applications are fully discussed in two published papers, which are included below.

Another discussion that was prompted during the involvement of the PRIDE system in disseminating the results of the HUPO Plasma Proteome Project (HPPP) pilot centered on what data types to make available. The issue of whether the proprietary binary data as recorded by the instrument should be made available in addition to the processed (text-based) peak lists typically used for the identification process was discussed. This discussion has been explained and a choice has been motivated in the paper included in subsection 2.4.1.3.

2.4.2. Publications

2.4.2.1. Initial publication in Proteomics

[59] Paraphrasing twelfth-century French philosopher Bernard of Chartres.

REGULAR ARTICLE

PRIDE: The proteomics identifications database

Lennart Martens[1], Henning Hermjakob[2], Philip Jones[2], Marcin Adamski[3], Chris Taylor[2], David States[3], Kris Gevaert[1], Joël Vandekerckhove[1] and Rolf Apweiler[2]

[1] Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium
[2] EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[3] Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

The advent of high-throughput proteomics has enabled the identification of ever increasing numbers of proteins. Correspondingly, the number of publications centered on these protein identifications has increased dramatically. With the first results of the HUPO Plasma Proteome Project being analyzed and many other large-scale proteomics projects about to disseminate their data, this trend is not likely to flatten out any time soon. However, the publication mechanism of these identified proteins has lagged behind in technical terms. Often very long lists of identifications are either published directly with the article, resulting in both a voluminous and rather tedious read, or are included on the publisher's website as supplementary information. In either case, these lists are typically only provided as portable document format documents with a custom-made layout, making it practically impossible for computer programs to interpret them, let alone efficiently query them. Here we propose the proteomics identifications (PRIDE) database (http://www.ebi.ac.uk/pride) as a means to finally turn publicly available data into publicly accessible data. PRIDE offers a web-based query interface, a user-friendly data upload facility, and a documented application programming interface for direct computational access. The complete PRIDE database, source code, data, and support tools are freely available for web access or download and local installation.

Received: October 11, 2004
Revised: January 25, 2005
Accepted: March 1, 2005

Keywords:
Bioinformatics / Databases / Protein identification

Correspondence: Dr. Lennart Martens, Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium
E-mail: lennart.martens@UGent.be
Fax: +32-9264-9484

Abbreviations: GPS, general proteomics standards; **PDF**, portable document format; **PPP**, Plasma Proteome Project; **PRIDE**, proteomics identifications database; **PSI**, proteomics standards initiative; **RDBMS**, relational database management system; **SQL**, structured query language; **UM**, University of Michigan; **W3C**, WWW Consortium; **XML**, extensible markup language; **XSL**, XML stylesheet language

1 Introduction

The field of proteomics has rapidly grown into one of the most active research areas in life sciences today. This growth is largely attributable to the availability of ever increasing amounts of gene and protein sequence information and the many technical improvements in the elaborate machinery used to identify proteins in complex mixtures, along with many novel techniques that reduce the complexities of analyte mixtures, allowing protein identification and characterization at an ever increasing

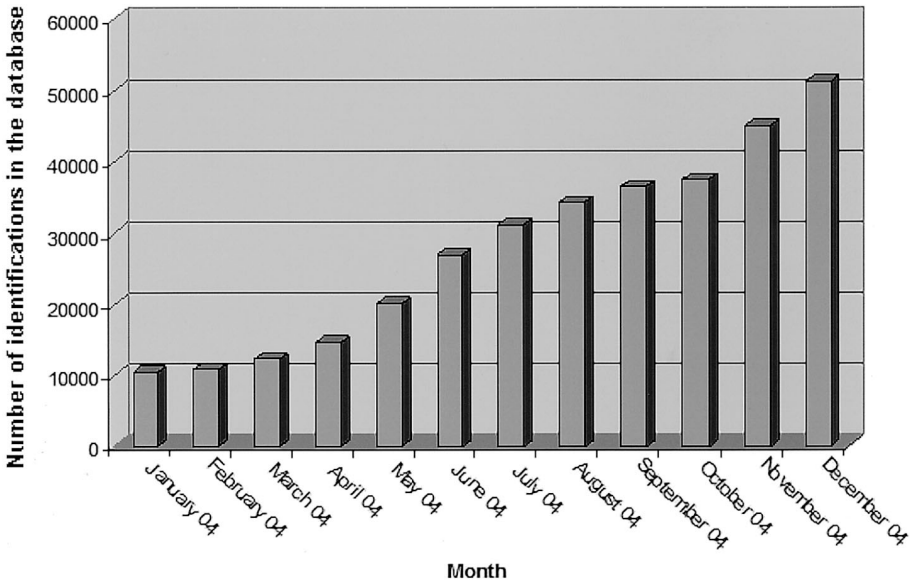


Figure 1. Illustration of the increase in the number of protein identifications over time. Number of identifications in the local database of the Department of Biochemistry at the University of Ghent throughout 2004 is shown. Source data originate from fragmentation spectra obtained from three different mass spectrometers: ESI-Q-TOF, ESI-IT, and MALDI-TOF/TOF. Note that the average increase since January 2004 amounts to 4000 identifications *per* month, often punctuated by particularly sharp increases when all three machines are fully operational in parallel. The vast majority of the identifications stems from experiments using the COFRADIC gel-free technology [2–4].

pace, as reviewed in [1]. To illustrate this further, the actual identification rate for a typical proteomics laboratory is given in Fig. 1.

Correspondingly, publication of protein identification data has been steadily on the rise over the past few years. The number of hits returned *per* year since 2000 for a PubMed query illustrates this in a simplistic, yet straightforward manner (Fig. 2). Together with the growing number of publications, the lists of identifications have grown considerably in size as well. Since these listings can easily contain thousands of peptides or hundreds of proteins, they are often published as supplementary information. In almost all cases this supplementary information consists of one or more portable document format (PDF) files detailing the identifications in a tabular format.

Yet even though PDF does an admirable job as a truly portable format and is therefore a natural choice for publishers, it is definitely not designed to convey structured information. Tables in PDF are notoriously difficult to extract and this problem is further exacerbated by the fact that nearly every author uses a different formatting for these tables.

Since this relative inaccessibility of proteomics data presents a considerable stumbling block on the way to making all these identifications really count for life sciences, the construction of a centralized, freely accessible repository for proteomics data is one of the primary requirements in proteomics today [5, 6]. In a single sentence: publicly *available* data needs to become publicly *accessible* data.

The pilot phase of the Plasma Proteome Project (PPP) [7], the first of the HUPO proteomics projects [8] to reach an important milestone, has been invaluable in achieving the ambitious goal of designing and implementing such a repository [9].

Indeed, the need for a centralized data repository was quickly realized during the initial planning phase for the PPP and this resulted in both a short-term and a long-term

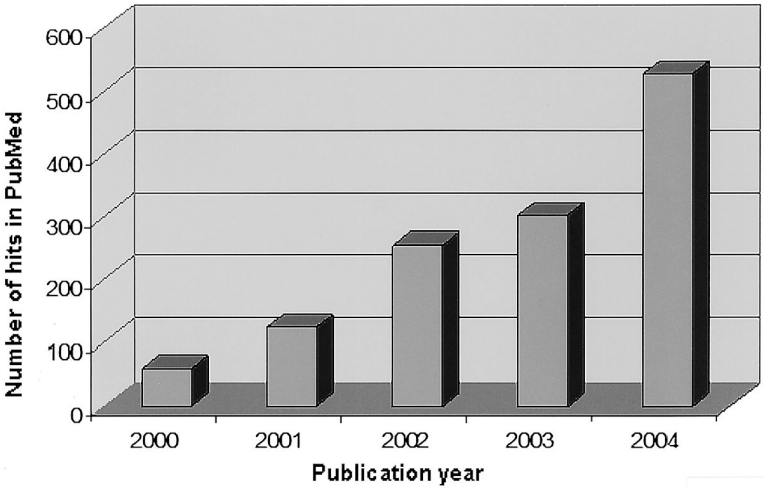


Figure 2. Illustration of the increase in protein identification papers over the past few years. Number of PubMed hits for each year since 2000 for the query "*proteom* AND proteins AND identified AND mass spectrometry*" are shown. Although this query is by no means exhaustive, it provides a meaningful sampling of the available literature.

approach to solving the data management problems. The short-term solution dealt with the immediate need for data storage and consisted of a relational database implemented using Microsoft Structured Query Language Server (MS-SQL). This database was constructed and continuously updated by Marcin Adamski from the core bioinformatics unit of David States at the University of Michigan (UM), Ann Arbor. This database actually had a two-fold objective: first of all, it served the vital purpose of centralizing the data produced in the PPP collaboration as it started to trickle (and later pour) in from the different labs, and second, it served as a test-bed for the construction of a centralized, project-independent database for protein identifications at the European Bioinformatics Institute (EBI). The aims of proteomics identifications database project are three-fold: developing an open source, publicly available set of tools to aid developers in implementing ms are three-fold: (1) providing a central repository for protein identification data, (2) building an efficient web-based interface for queries and data submission, and (3) developing an open source, publicity available set of tools to aid developers in inplementing custom analysis tools.

The MS-SQL database constructed at UM proved to be an excellent source of inspiration for the PRIDE data model as it had been refined throughout the PPP in order to contain detailed proteomics data from many different collaborating laboratories across the globe. The design of PRIDE and the functionality of its web interface will be discussed next, along with future prospects for the data model as proteomics standards evolve.

2 Materials and methods

The PRIDE project was completely developed in the Java 2 programming language (Sun Microsystems) using the Java™ Development Kit (JDK) 1.4 from Sun Microsystems (http://www.java.com/en/download/manual.jsp) as well as the Java™ 2 Enterprise Edition (J2EE) extensions from Sun Microsystems (http://java.sun.com/j2ee/index.jsp) for the web development.

PRIDE makes use of many open source software tools, components, and libraries. Object-relational bridge (OJB) (http://db.apache.org/ojb) takes care of the declarative object-relational mapping, Tomcat (http://jakarta.apache.org/tomcat) functions as web server and servlet engine, Log4J (http://logging.apache.org/log4j) as the logging framework, and Maven (http://maven.apache.org) as the project management tool. All of the above were obtained from the Apache Software Foundation (http://www.apache.org). Extensible markup language (XML) parsing and writing relies on the XML pull parser (XPP) libraries (http://www.extreme.indiana.edu/xgws/xsoap/xpp). Unit testing was performed using the JUnit framework (http://www.junit.org). During development, the relational database management system (RDBMS) employed was MySQL (http://www.mysql.com) and for the final prototyping and production version PRIDE was ported to Oracle (http://www.oracle.com).

3 Results and discussion

3.1 PRIDE as a set of components

The PRIDE project consists of a number of distinct parts, which are summarized in Fig. 3. The XML format represents the basic data structure, whereas the relational database implementation is just one of the possible renderings of the hierarchical XML format in a relational schema. The PRIDE core libraries contain an object model of the PRIDE data structure and allow the programmer to interact seamlessly and effortlessly with the PRIDE XML format and reference database implementation. The PRIDE web libraries provide a web-based view on an underlying reference database and use the PRIDE core libraries for data access. Query results from the web can be sent in PRIDE XML format or in HTML after XML stylesheet language (XSL) transformation of the XML.

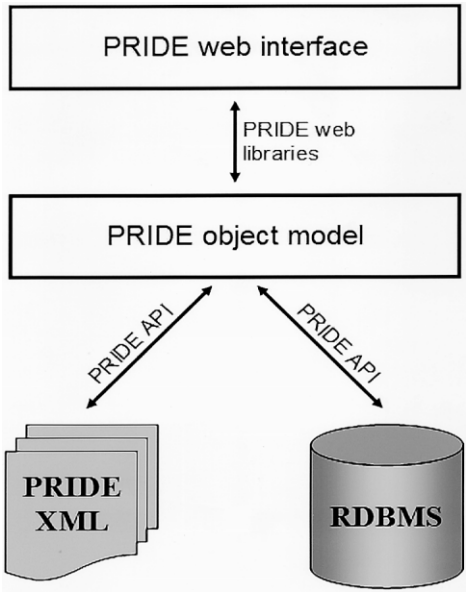


Figure 3. PRIDE components. PRIDE project consists of a number of separate components which are outlined here. PRIDE XML format is the basic data structure. RDBMS implementation is a possible rendering of the hierarchical XML format into a relational schema. PRIDE core libraries constitute the basic object model representations of the PRIDE data structure, as well as I/O objects that allow easy interaction with both the reference database implementation and XML format. PRIDE web libraries have been built to allow web-based submission and access to PRIDE. Web libraries use the PRIDE core libraries for data access and processing and can report query results in PRIDE XML as well as in XSL-transformed HTML.

3.2 XML data structure

XML is a standard text format developed by the WWW Consortium (W3C, http://www.w3c.org) that has quickly become a very popular and widely applied means of storing data as well as exchanging them. XML is a hierarchical (tree-like) structure, which fits well with typical proteomics experiments and can easily be validated. XML documents can also readily be extended, which allows them to retain a rather large degree of flexibility. For these reasons, XML was chosen to form the basic data structure for the PRIDE project rather than a relational database structure.

3.3 Relational database implementation

Relational databases provide highly efficient storage of structured data and can readily be optimized for extremely fast retrieval of data based on queries. These queries can be fed to the database through the use of a standardized interface: SQL.

Contrary to the traditional approach, which relies on a relational schema as the basic data structure, PRIDE instead builds upon the XML schema for reasons discussed above. Since XML is hierarchical in nature, and a database is relational, the mapping of an XML schema to a database schema is not straightforward. In fact, many different relational approaches can model exactly the same hierarchical schema. Therefore, the reference implementation provided by PRIDE is just one of the possible forms this database might take. The strength of the XML schema-based structure is that, depending on specific needs, other relational implementations can be created by third parties that emphasize or optimize different aspects of the data stored. As such, PRIDE can be molded to take many different queryable forms, each with distinct strengths and weaknesses.

3.4 The PRIDE data format

In PRIDE, one or more experiments are contained in the root tag "ExperimentCollection". The ExperimentCollection simply groups together one or more "Experiment" tags, which are the top-level tags for individual results. As such, a submitter is likely to submit a collection containing a single experiment, unless the data are extensive or varied enough to warrant the creation of multiple, distinct Experiment elements within PRIDE. The downloadable flat file on the other hand, will hold an ExperimentCollection root consisting of all experiments that constitute the PRIDE database at its release time. The ExperimentCollection structure will enable easy splitting of the download into multiple files when the PRIDE download file eventually becomes very large.

The top-level structure of an experiment, schematically represented and exemplified in Fig. 4, consists of seven conceptually distinct parts, which will be summarized next.

The first of these is the experiment accession number. This number is assigned after successful submission of an experiment and provides a unique pointer to all the associated data. The experiment accession number would be the data element of choice for inclusion in papers as the PRIDE reference because of its conciseness and since interested readers can easily use it to quickly retrieve all relevant data from the PRIDE web interface.

The second part contains meta-data about the experiment: a descriptive title, contact person and/or address, a short label, a description, and finally location information. The contact person and location information are meant to be complementary, *i.e.*, to have geographic and laboratory information in the location field, and contact person information in the contact information. Typically, one would expect the "contact person" field to contain the e-mail address of the corresponding author of a publication.

The third element concerns the sample studied. It consists of a description field and an attribute list. The structure and usage of the latter is discussed in more detail below.

Protocol information constitutes the fourth part of an experiment. Apart from a description and attribute list, it also holds one or more sections about the mass spectrometer(s) used. This latter section contains manufacturer, model, source, and analyzer information which can be further supplemented through an attribute list.

The fifth part details the information derived from the mass spectrometer. This section holds the MS coefficient (*e.g.*, MS^2, MS^3), peak lists, optional raw data references, comments, and an attribute list.

The most intricate subsection of an experiment is the sixth part and deals with the identifications obtained from the data specified in part five. Identifications have been split in two different types: 2-D PAGE-based identifications and nongel-based identifications. A schematic representation of the shared and specific elements for both of these subtypes is shown in Fig. 5. The shared elements are wrapped up in an abstract ancestor element called "IdentificationType". Note that the additional information for 2-D PAGE-based identifications centers on protein-related data gathered during the gel-separation phase, whereas the gel-free identifications typically require more information about the effective identification score and threshold (if available). This has been done to accommodate more stringent standards for identifications, as discussed in recent publications [10–12].

Finally, the seventh part is not restricted to the experiment level but can be found in many of the smaller branches as well. This is the "AttributeList" which represents a list of attributes, to be keyed from controlled vocabularies, allowing an extremely flexible way of integrating additional information into the core schema without sacrificing the structure of the whole. In fact, the PRIDE schema presents a *minimum minimorum* of information about protein identifications in present day proteomics. Many additional pieces of information (*e.g.*, from cone voltages and temperatures on ESI-type ion sources to the specific search parameters used for

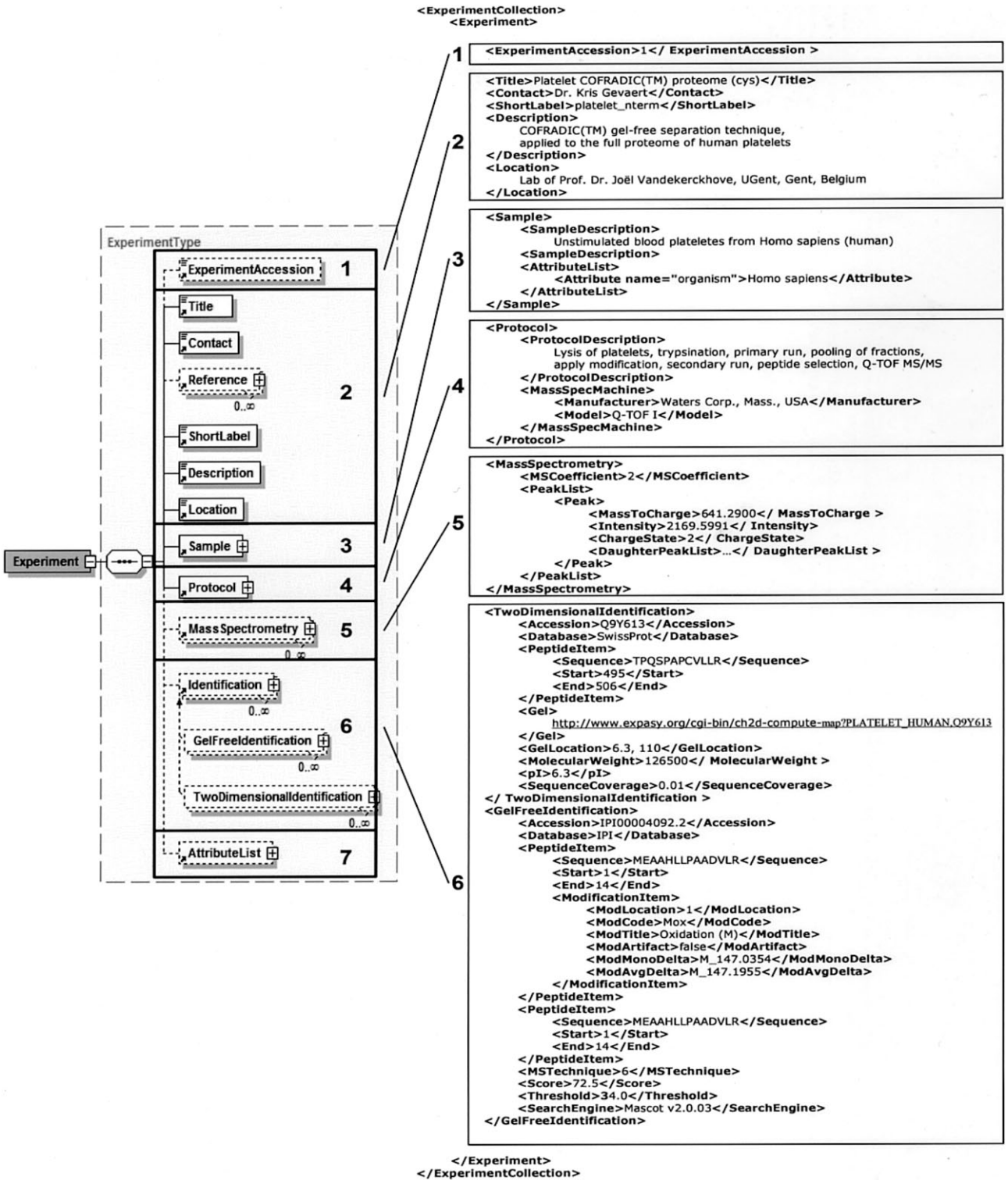


Figure 4. Top-level view of the PRIDE XML data structure and example document. Boxed and numbered areas represent conceptually distinct parts of the Experiment node top-level structure (see text for details). Optional elements are indicated by dashed boxes and multiplicity (if applicable) is shown below the box on the right. Note that an abbreviation has been introduced in the "PeakList" element due to space constraints. Recursive "DaughterPeakList" element is symbolically filled out by "...". An example of a simple AttributeList element (number 7) is represented in the "Sample" element (number 3). Please also note that the presented document is meant to provide an example only. As such, the occurrence of a "TwoDimensionalIdentification" in what is described as a "gel-free separation technique" has only demonstrative purposes.

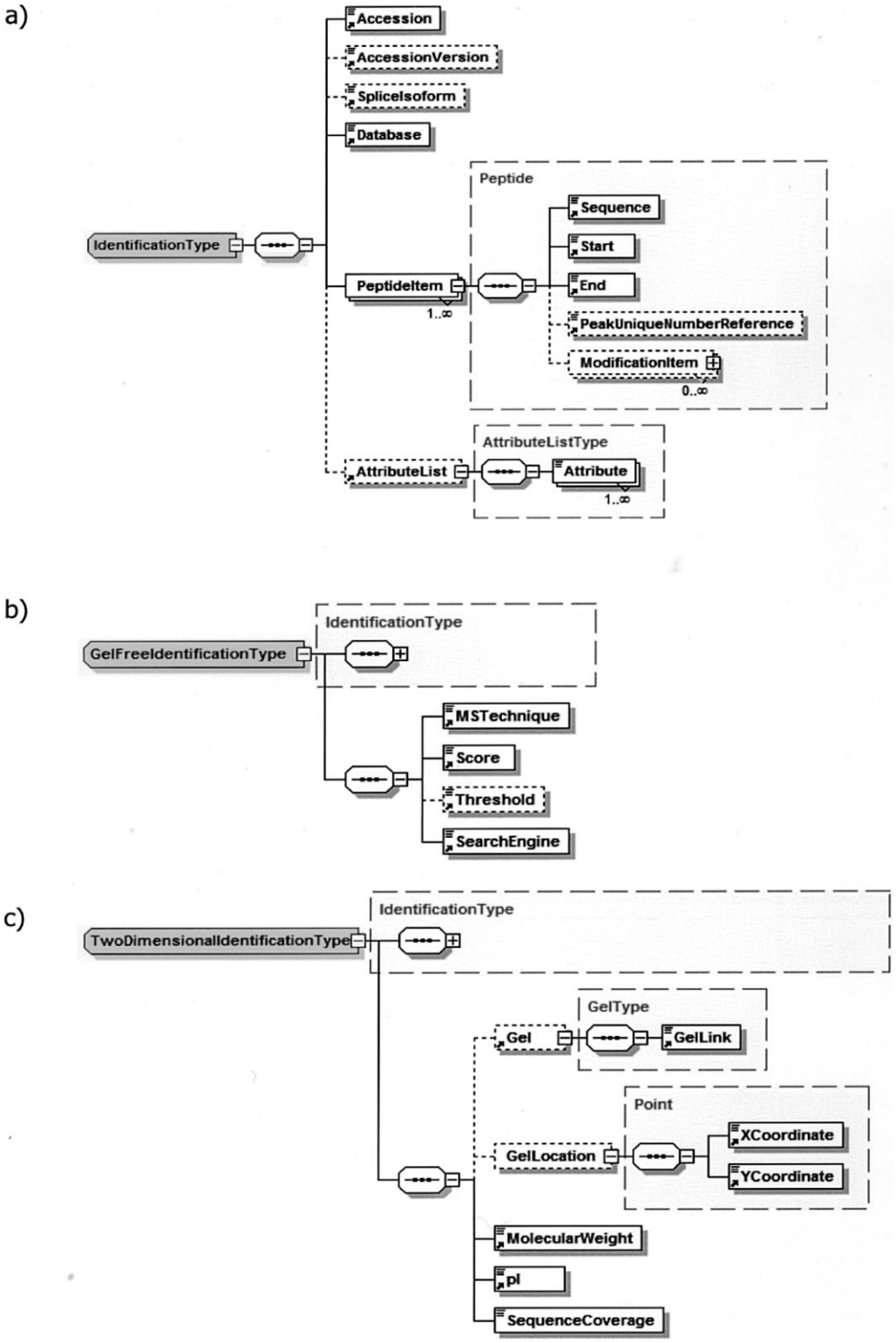


Figure 5. Detailed view of the identification data structures. Abstract ancestor element IdentificationType, shown in (a) contains the shared data elements for the two implementing forms, "GelFreeIdentificationType" (b) and "TwoDimensionalIdentificationType" (c). Both have specific properties which are displayed for each. Note that the properties that were inherited from IdentificationType have been collapsed for clarity in (b) and (c). Optional elements are indicated by dashed boxes and multiplicity (if applicable) is shown below the box on the right.

querying a sequence database with a fragmentation spectrum) that are gathered during the identification process are extremely useful to other researchers, yet few people actively gather and store this information in a structured way. Therefore, the inclusion of this information cannot now be mandatory, nor is it possible to mold it to a defined structure. Indeed, the field of proteomics is still evolving quite rapidly, and allowing for this kind of semistructured data makes PRIDE flexible enough to accommodate future requirements without a major overhaul. In the long term, it is conceivable that some of these attributes become standard elements, whereas others will become obsolete.

PRIDE will in fact be gradually extended to embrace the proteomics standards initiative-general proteomics standards (PSI-GPS) [13] as they become available, thus shaping the PRIDE format into an implementation of this broader format. Most notably, the mzData format for storage of mass-spectrometer derived information is quickly reaching maturity, and as soon as the controlled vocabularies for this format are released (expected by spring 2005), PRIDE will incorporate this format. The mzIdent format, meant to capture the identifications that result from searches based on MS data, is in a more primitive stage at this point, but PRIDE will adopt this standard upon availability as well. The success of previous PSI standard formats [14] has led us to committing ourselves to their GPS standards, yet other proposed standards for data interchange formats which have been published in peer-reviewed literature [15, 16] will also be accommodated by automated conversion in the near future.

3.5 Comparison between the PRIDE data format and the PPP database at UM

Even though the PRIDE database draws in part upon the PPP structure devised at UM, both models do not overlap in full. This is mainly due to the slightly different focus of the respective databases. PRIDE, being developed as a generic repository for proteomics data, necessarily lacks some of the very detailed structures present in a single-purpose database such as the PPP database at UM. Specifically, the PPP database provides more structured detail both at the level of the protocol description as well as the identification process (including full details about the searches performed). This additional level of detail in the PPP database enables the collaboration to compare the findings across similar yet slightly different plasma samples, across technologies and platforms used, and across identification algorithms applied to the data.

Although this information is not present in PRIDE as structured data (*i.e.*, there are no database columns or XML tags with corresponding names), the ability to cope with this data is inherently present through the use of attribute lists. Coupled to controlled vocabularies, which can be both particular to as well as shared across experiments, these attribute lists enable storage of any desirable additional level of detail at several crucial points in the PRIDE data structure.

3.6 The PRIDE web interface

PRIDE presents a default web interface that provides three areas of functionality: the ability to search and query the PRIDE database, the facility to register as a data submitter or collaborator, and the ability to submit data to PRIDE.

Five types of queries are supported in the current release (Fig. 6). Queries on experiment title and accession number are particularly useful when the user wants to view the full list of identifications for an (published) experiment. Additionally, queries can also be performed on a text fragment from a reference, enabling users to obtain the identifications associated with a certain publication or author. Querying the database by protein accession number lists all known identifications of the specified protein across experiments, together with detailed identification information such as the peptides identified and their modifications. Finally, PRIDE can be searched by sample name. This allows the user to see all proteins identified in a certain tissue, cell type, or organism, again across all experiments and with full details.

The PRIDE web interface can be configured to return HTML formatted results as well as XML formatted results. As such, it caters for both human readers (HTML format) and machine readers (XML format). The latter allows users to write scripts that perform an off-line meta-analysis on the results of PRIDE queries in an efficient way.

The ability to register as a data submitter or collaborator has been implemented with several goals in mind. First, restricting data submission to registered users coupled with a simple level of data curation will help to avoid spurious data being uploaded into PRIDE. The system allows the creation of collaborations, such that submitted data can be kept private and shared only amongst collaborators until such a time as they wish to make their data public. This functionality allows PRIDE to be used as a tool for collaboration and data sharing within a consortium as well as serving as a final repository for published data. The same applies to PRIDE as a system for peer-reviewing data that has been submitted for publication, since data can be privately shared between author(s), the journal, and the reviewers. Obviously it is also possible for a data submitter to declare their data publicly at the point of submission, excluding it from any of the restrictions described above.

4 Concluding remarks

The PRIDE project has resulted in the construction of a unique combination of tools, standards, and infrastructure that for the first time enable the construction of a truly global, centralized proteomics data repository. The highly modular design makes PRIDE flexible enough to be adapted by third parties to create more or less differing mirrors with new or specialized views on the same data.

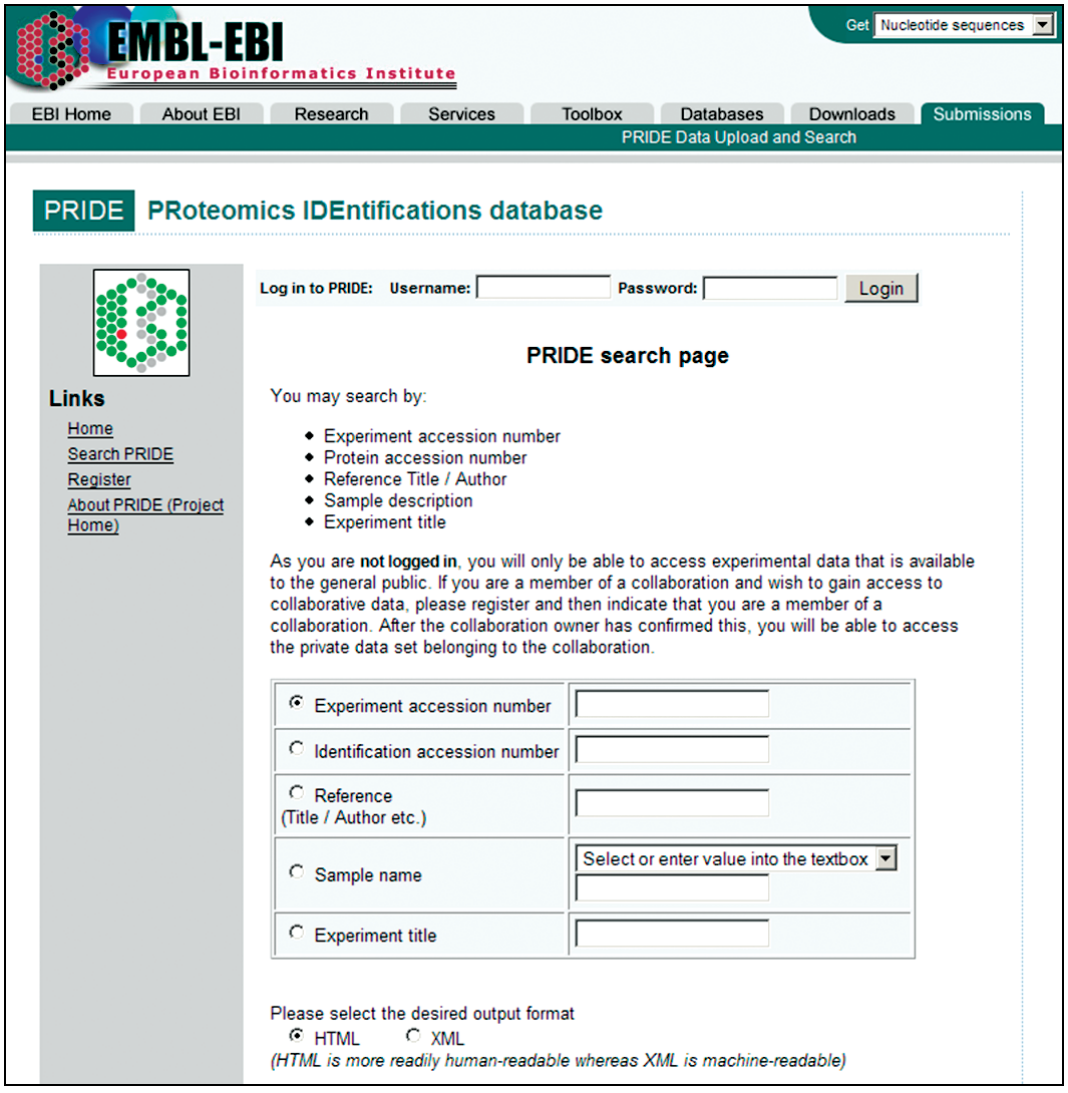


Figure 6. PRIDE web interface for queries. Three types of query are supported (1) by experiment accession number, (2) by protein accession number, (3) by reference text fragment, (4) by sample, or (5) by experiment title.

Indeed, as PRIDE starts to gather data, we expect new and unexpected uses of the publicly available data to come up. Statisticians might seize the opportunity to constructively contribute to the way database search software functions or could enhance the procedures used to distinguish between true identifications and false positives. Biologists can mine the data in search for new research targets and software developers can come up with new ways to store, visualize, and query the large amounts of data that will accumulate over time.

It is our conviction that PRIDE will be an important milestone in the evolution of the field of proteomics and it is our hope that it will become the highly active hub of proteomics data that we designed it to be.

The PRIDE database can be accessed on-line at http://www.ebi.ac.uk/pride.

The authors would like to extend their gratitude to Gilbert S. Omenn for his support of the project and for the many informative discussions which have contributed to this paper. L.M. would like to thank Samuel Kerrien, Mark Rijnbeek, and Kai Runte for their invaluable comments, suggestions, and contributions to the development of the PRIDE code base, reference relational database implementation, and XML schema. The PRIDE project was funded in part through the European Commission Programme "Quality of Life", Marie Curie Training Site Fellowship, Contract number: QLRI-1999-50595. Parts of the data used in this paper were generated in the context of the IWT-GBOU-research initiative (Project number 20204) of the Flanders Institute of Science and Technology (IWT). K.G. is a Postdoctoral Fellow and L.M. a Research Assistant of the Fund for Scientific Research-Flanders (Belgium) (F.W.O. Vlaanderen).

5 References

[1] Zhang, H., Yan, W., Aebersold, R., *Curr. Opin. Chem. Biol.* 2004, *8*, 66–75.

[2] Gevaert, K., Van Damme, J., Goethals, M., Thomas, G. R., Hoorelbeke, B., Demal, H., Martens, L. , *Mol. Cell. Proteomics* 2002, *1*, 896–903.

[3] Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G. R., Vandekerckhove, J., *Nat. Biotechnol.* 2003, *21*, 566–569.

[4] Gevaert, K., Ghesquière, B., Staes, A., Martens, L., Van Damme, J., Thomas, G. R., Vandekerckhove, J., *Proteomics* 2004, *4*, 897–908.

[5] Prince, J. T., Carlson, M. W., Wang, R., Lu, P., Marcotte, E. M., *Nat. Biotechnol.* 2004, *22*, 471–472.

[6] Rohlff, C., *Expert Rev. Proteomics* 2004, *1*, 267–274.

[7] Omenn, G. S., *Proteomics* 2004, *4*, 1235–1240.

[8] Hanash, S., Celis, J. E., *Mol. Cell. Proteomics* 2002, *1*, 413–414.

[9] Adamski, M., Blackwell, T. W., Menon, R., Martens, L., Hermjakob, H., Taylor, C. F., Omenn, G., States, D., *Proteomics* 2005, *5*, this issue.

[10] Carr, S. A., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., Nesvizhskii, A., *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[11] Veenstra, T. D., Conrads, T. P., Issaq, H. J., *Electrophoresis* 2004, *25*, 1278–1279.

[12] Nesvizhskii, A. I., Aebersold, R., *Drug Discov. Today* 2004, *9*, 173–181.

[13] Orchard, S., Taylor, C. F., Hermjakob, H., Zhu, W., Julian, R. K. Jr., Apweiler, R., *Proteomics* 2004, *4*, 2363–2365.

[14] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Saliwinski, L., Ceol, A., Moore, S. *et al.*, *Nat. Biotechnol.* 2004, *22*, 177–183.

[15] McDonald, W. H., Tabb, D. L., Sadygov, R. G., MacCoss, M. J., Venable, J., Graumann, J., Johnson, J. R. *et al.*, *Rapid Commun. Mass Spectrom.* 2004, *18*, 2162–2168.

[16] Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B. *et al.*, *Nat. Biotechnol.* 2004, *22*, 1459–1466.

2.4.2.2. PRIDE 2.0 publication in Nucleic Acids Research

Nucleic Acids Research, 2006, Vol. 34, Database issue ***D659–D663***
doi:10.1093/nar/gkj138

PRIDE: a public repository of protein and peptide identifications for the proteomics community

Philip Jones[1,*], Richard G. Côté[1], Lennart Martens[2], Antony F. Quinn[1], Chris F. Taylor[1], William Derache[1], Henning Hermjakob[1] and Rolf Apweiler[1]

[1]EMBL Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and [2]Department of Medical Protein Research, Flanders Interuniversity Institute for Biotechnology, Faculty of Medicine and Health Sciences, Ghent University, Rommelaere Institute, Building D, A. Baertsoenkaai 3, B-9000 Ghent, Belgium

Received August 15, 2005; Revised and Accepted October 27, 2005

ABSTRACT

PRIDE, the 'PRoteomics IDEntifications database' (http://www.ebi.ac.uk/pride) is a database of protein and peptide identifications that have been described in the scientific literature. These identifications will typically be from specific species, tissues and sub-cellular locations, perhaps under specific disease conditions. Any post-translational modifications that have been identified on individual peptides can be described. These identifications may be annotated with supporting mass spectra. At the time of writing, PRIDE includes the full set of identifications as submitted by individual laboratories participating in the HUPO Plasma Proteome Project and a profile of the human platelet proteome submitted by the University of Ghent in Belgium. By late 2005 PRIDE is expected to contain the identifications and spectra generated by the HUPO Brain Proteome Project. Proteomics laboratories are encouraged to submit their identifications and spectra to PRIDE to support their manuscript submissions to proteomics journals. Data can be submitted in PRIDE XML format if identifications are included or mzData format if the submitter is depositing mass spectra without identifications. PRIDE is a web application, so submission, searching and data retrieval can all be performed using an internet browser. PRIDE can be searched by experiment accession number, protein accession number, literature reference and sample parameters including species, tissue, sub-cellular location and disease state. Data can be retrieved as machine-readable PRIDE or mzData XML (the latter for mass spectra without identifications), or as human-readable HTML.

INTRODUCTION

The vast quantity of data associated with a single proteomics experiment can become problematic at the point of publishing the results. Laboratories tend to publish their work in an appropriate journal with perhaps a PDF document listing the proteins described. If space allows, the individual peptide sequences may be included but there is little possibility of including details of the mass spectra in this format. This clearly creates difficulties while attempting to reproduce the work of a laboratory to confirm their results.

Fortunately, the community has recognized and is tackling this problem through the formation of groups concerned with the development of standards for the capture and sharing of proteomics data. One such group is the HUPO Proteomics Standards Initiative (PSI) (1) who are in the process of developing standards tackling several aspects of proteomics, including ontologies of proteomics related terms, XML schemata and minimal reporting guidelines.

The Proteomics Identifications database (PRIDE), previously described by Martens *et al*. (2) is a PSI compliant public repository for proteomics identifications to which any proteomics laboratory is welcome to submit data. It is envisaged, but not mandated, that any such submission would normally be in the context of the corresponding submission of a manuscript to a journal describing the identifications submitted to PRIDE. As such, PRIDE aims to become the proteomics equivalent of the ArrayExpress database (3) used to capture microarray experiment data in support of journal publications.

PRIDE is not alone in this endeavor. Several other publicly available databases exist for the purpose of capturing and disseminating proteomics data from mass spectrometry. Such databases include the Global Proteome Machine Database (gpmDB) (4), The Institute for Systems Biology's PeptideAtlas (5) and the University of Texas' Open Proteomics Database (opd) (6). Currently in progress is the development of a collaborative agreement to exchange data between

*To whom correspondence should be addressed. Tel: +44 1223 492610; Fax: +44 1223 494468; Email: pjones@ebi.ac.uk

these and other emerging proteomics data repositories, including PRIDE.

DATABASE DESCRIPTION

What is the scope of PRIDE?

PRIDE can store

(i) The title and description of the experiment, together with contact details of the submitter.
(ii) Literature references.
(iii) Protein identifications by accession number supported by a corresponding list of one or more peptide identifications.
(iv) For each peptide identified, the sequence and coordinates of the peptide within the protein that it provides evidence for. Optionally, a reference to any submitted mass spectra that form the evidence for the peptide identification.
(v) Any post-translational modifications (natural or artefactual) coordinated in relation to the specific peptide that they have been found upon.
(vi) A description of the sample under analysis, including but not limited to the species of origin, tissue, sub-cellular location (if appropriate), disease state and any other relevant annotation.
(vii) A description of the instrumentation used to perform the analysis, including mass spectrometer source, analysers and detector, instrument settings and software settings used in data processing to generate peak lists.
(viii) Processed peak lists supporting the identifications in PRIDE in the versatile PSI mzData format.

PRIDE version 2.0, the release of PRIDE available at the time of writing, makes use of HUPO PSI deliverables such as the mzData XML schema (7) for capturing the settings and output from mass spectrometry work flow, including items vi–viii listed earlier. At present, the PRIDE XML schema encompasses the mzData schema with additional elements to allow protein and peptide identifications and post-translational modifications to be captured. It is envisaged that the analysisXML XML schema will be incorporated into PRIDE following its first release as a finalized schema, expected by early 2006, replacing large parts of the custom schema currently present in PRIDE.

Datasets currently available in PRIDE

A significant dataset that is publicly available from PRIDE at the time of writing is the set of protein and peptide identifications from the individual laboratories involved in the HUPO Plasma Proteome Project (8). This project was in part responsible for the requirements statement that initiated the PRIDE project.

Another publicly available dataset in PRIDE is a profile of the human platelet proteome (9) submitted by the Department of Medical Protein Research, Ghent University. This department is also scheduled to contribute a substantial dataset identifying proteolytic cleavage by caspases in apoptotic Jurkat T-cells (10) as well as a large set of spectra used to evaluate spectrum quality filtering software (11).

A dataset of protein and peptide identifications describing the organelle proteome of the secretory pathway is currently held as private data in PRIDE but is expected to be publicly available following publication of the related manuscript. At present this dataset can only be viewed by prior permission of the submitters.

It is expected that by the end of 2005 PRIDE will also contain the protein and peptide identifications and related mass spectra from the HUPO Brain Proteome Project (12) as a publicly available dataset.

Submission and retrieval of data

Data can be both submitted to and retrieved from PRIDE through a web interface, using either the PRIDE XML schema, which embeds mzData as a sub-element to allow inclusion of details of the spectra, or using the mzData XML schema, in which case all identifications will be omitted.

Data can also be viewed as a human-readable HTML table illustrated in Figure 1.

Figure 2 illustrates the search page. Queries can include experiment identifier, protein accession or identifier, literature references and sample parameters, including species, tissue, sub-cellular location and disease. The search results include all the experiments that match the query, together with options of how the data should be presented.

Data security in PRIDE: PRIDE as a tool for journal review

Data submitted to PRIDE is marked as public or private. Private data can be shared through a collaborative mechanism that allows individuals to apply to join a collaboration, their application then being confirmed or rejected by the creator of the collaboration. As well as allowing collaborating laboratories to share their data, this mechanism can also be used to allow manuscript reviewers to access the corresponding PRIDE entry in a confidential manner on a neutral site.

Use of controlled vocabularies and ontologies in PRIDE

By extending the mechanism designed for the mzData XML schema, PRIDE makes extensive use of external controlled vocabularies and ontologies (hereafter 'CVs') to annotate entries. The use of CVs ensures that queries for particular terms will capture all of the relevant data without omission due to differences in terminology. As a spin-off of the PRIDE development program, a SOAP web service to allow external CVs to be queried in an intelligent manner has been developed at the EBI, initially for use by PRIDE (http://www.ebi.ac.uk/ontology-lookup/). This service allows queries to take advantage of the hierarchical nature of ontologies. For example, if a user requests all protein identifications found in *pancreas*, the relevant Medical Subject Headings (MeSH) term will be looked up in the ontology web service and PRIDE will be queried for entries relating to the MeSH term 'Pancreas' as well as all child terms, currently in this case including 'Islets of Langerhans', 'Pancreas, Exocrine' and 'Pancreatic Ducts'. This mechanism assists the user by retrieving all the relevant data without the need to have a detailed knowledge of the terms involved.

CVs and ontologies suggested for use in PRIDE include MeSH (13) for animal anatomy and disease states; Gene

PRIDE Experiment Collection Version 2.0

Experiment 1: COFRADIC methionine proteome of unstimulated human blood platelets

Experiment: (top ↑)	**Description:** COFRADIC methionine proteome of unstimulated human blood platelets **Short Title:** Platelets MetOx **Accession:** 1
References:	Martens, L., Van Damme, P., Van Damme, J., Staes, A., Timmerman, E., Ghesqui?re, B., Thomas, G.R., Vandekerckhove, J., Gevaert, K., Proteomics, in press
Protocol:	**Name:** methionine oxidation induces a chromatographic shift on a diagonal RP-HPLC system.

Identifications

Accession	Splice Isoform	Database	Score	Threshold	Search Engine	Additional Information (Source / Name / Value)
IPI00295313		IPI human 2.31	52.0	35.0	Mascot 2.0.03	
IPI00017340		IPI human 2.31	56.0	35.0	Mascot 2.0.03	
IPI00026128		IPI human 2.31	80.6667	27.3333	Mascot 2.0.03	
IPI00031169		IPI human 2.31	48.0	33.5	Mascot 2.0.03	
IPI00291262		IPI human 2.31	60.6667	41.0	Mascot 2.0.03	
IPI00395553		IPI human 2.31	48.0	35.0	Mascot 2.0.03	
IPI00328748		IPI human 2.31	43.0	37.0	Mascot 2.0.03	
IPI00027497		IPI human 2.31	48.0	35.3333	Mascot 2.0.03	

mzData

Information:	**Version:**	1.05
	Accession:	1
Sample:	**Name:**	unstimulated human platelets
	Description:	
	Additional:	Source: NEWT, Name: Homo Sapiens, Value:
		Source: MeSH, Name: blood_platelets, Value:
Source File:		
Contact:	**Name:**	Kris Gevaert
	Institution:	Ghent University, Dept. of Medical Protein Research
	Contact information:	kris.gevaert@UGent.be
Instrument:	**Name:**	Micromass UK Limited, Cheshire, UK Q-TOF I
	Source:	Source: PSI, Name: IonizationType, Value: ESI
		Source: User, Name: comment, Value: Fragmentation time per ms/ms spectrum was 8 sec
	Analyzer #1:	Source: PSI, Name: AnalyzerType, Value: Q-TOF
	Detector:	Source: User, Name: No Detector Componen..., Value:
	Additional Information:	Source: PSI, Name: Vendor, Value: Micromass UK Limited, Cheshire, UK
		Source: PSI, Name: Model, Value: Q-TOF I

Figure 1. An example of PRIDE data in tabulated HTML format.

Ontology (GO) (14) for sub-cellular location; NEWT (15) for taxonomy, which is a superset of the NCBI taxonomy (16); the mass spectrometry ontology being developed by the HUPO PSI; RESID for naturally occurring post-translational modifications (17) and UNIMOD for protein modifications encountered in mass spectrometry experiments (18).

A PRIDE CV has been created for cases where existing CVs do not include a term required to annotate data in PRIDE.

The use of CVs and ontologies will allow the annotation of certain specific experimental results such as peptide retention times for LC-MS experiments or protein quantitation information for quantitive or differential proteomics experiments.

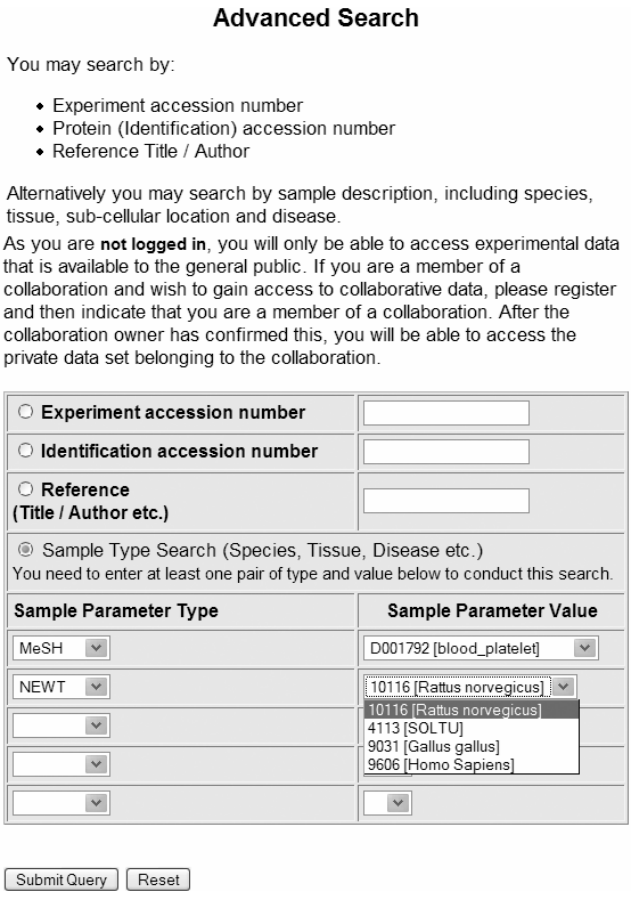


Figure 2. The PRIDE Advanced Search form.

Where the required CV terms do not exist already, PRIDE can accommodate these data elements through the use of user parameters.

PRIDE is an open-source software development project

Care has been taken throughout the development of PRIDE to ensure that all system components are open-source and freely available. PRIDE is written in Java and made available under the open-source Apache license. All the source code are freely available from the CVS repository (http://sourceforge.net/cvs/?group_id=122040). PRIDE uses the open-source Object-Relational Bridge (OJB) (http://db.apache.org/ojb/) API for database connectivity. As a consequence, PRIDE can easily be adapted to run on any SQL-based relational database management system. Configuration files exist for both Oracle (http://www.oracle.com) and MySQL (http://www.mysql.com/).

DISCUSSION

Here we consider possible applications of PRIDE from the perspective of the typical proteomics researcher. PRIDE offers the user several useful query opportunities including:

- Retrieving all proteomics experiments in which a particular protein of interest has been observed.
- Downloading proteome datasets of interest in a standard format for further local analysis.
- Retrieving the complete list of protein identifications (and the specific peptides found) for a given publication.
- Using the links provided via PRIDE to further explore the proteins identified.
- Comparison of one's own results with previous findings to quickly determine overlap as well as potentially novel findings.
- Planning of one's experiments: finding experimental protocols that have already been applied successfully to analyse your sample or even protein of interest.
- Re-analysing previously published results using your own techniques.
- Obtaining test sets for training or trying out novel algorithms (e.g. algorithms for protein or peptide identification).
- Retrieving typical base line proteomes for specific tissues and species.
- Allowing journal appointed reviewers to analyse the details of the identifications and potentially the supporting spectra as part of their review in a standardized manner.

FUTURE DEVELOPMENTS

The developers of PRIDE recognize that the system has room to evolve in several important aspects.

It is important for the future of PRIDE to keep apace with developments in the HUPO PSI. One important development of this initiative is the analysisXML XML schema, designed to hold details of protein and peptide identifications and

post-translational modifications, together with cross references to the relevant mzData entries describing spectra. It is intended that analysisXML will be fully supported by PRIDE for import and export, without modification or data loss, as soon as possible after the first stable release of the new analysisXML XML schema.

Submitters of identifications to PRIDE will naturally make use of their favored protein sequence database against which to search their spectra. Consequently, PRIDE will quickly fill with protein accessions and IDs from disparate protein sequence databases. An important short-term goal of the PRIDE project is to map all identifications to the UniProt database (19), including cross references to as many other protein databases as possible. This work will borrow heavily from the IntAct project (20), both in terms of code base and procedures for automatic and human curation.

A long-term goal of the PRIDE project is to provide an automated program of regular re-analysis of mass spectra deposited in PRIDE using the most up-to-date protein sequence databases and available open-source search algorithms such as X!Tandem (http://www.thegpm.org/TANDEM/) (21). The submitter's original identifications would continue to be available as described in the corresponding manuscript.

The EBI has developed a Distributed Annotation Server (DAS) (22) service for PRIDE (http://www.ebi.ac.uk/das-srv/pride/das/) using the BioJava Dazzle servlet (http://www.biojava.org/dazzle/). This service is publicly available and can be used to enable DAS clients such as Dasty (23), designed for visualizing protein sequence and annotation, to display identifying peptides for the protein specified in the DAS request.

ACKNOWLEDGEMENTS

PRIDE is supported through BBSRC iSPIDER and HUPO Plasma Proteome Project funding as well as a EU Marie Curie fellowship. L.M. is a research assistant of the Fund for Scientific Research, Flanders (Belgium) (F.W.O. Vlaanderen). L.M. would like to thank Prof Dr Kris Gevaert and Prof Dr Joël Vandekerckhove for their support. Funding to pay the Open Access publication charges for this article was provided by BBSRC iSPIDER and HUPO.

Conflict of interest statement. None declared.

REFERENCES

1. Orchard,S., Hermjakob,H. and Apweiler,R. (2003) The proteomics standards initiative. *Proteomics*, **3**, 1374–1376.
2. Martens,L., Hermjakob,H., Jones,P., Adamski,M., Taylor,C., States,D., Gevaert,K., Vandekerckhove,J. and Apweiler,R. (2005) PRIDE: the proteomics identifications database. *Proteomics*, **5**, 3537–3545.
3. Parkinson,H., Sarkans,U., Shojatalab,M., Abeygunawardena,N., Contrino,S., Coulson,R., Farne,A., Lara,G.G., Holloway,E., Kapushesky,M. *et al*. (2005) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*., **33**, D553–D555.
4. Craig,R., Cortens,J.P. and Beavis,R.C. (2004) Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res*., **3**, 1234–1242.
5. Deutsch,E.W., Eng,J.K., Zhang,H., King,N.L., Nesvizhskii,A.I., Lin,B., Lee,H., Yi,E.C., Ossola,R. and Aebersold,R. (2005) Human Plasma PeptideAtlas. *Proteomics*, **5**, 3497–3500.
6. 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.
7. Orchard,S., Hermjakob,H., Taylor,C.F., Potthast,F., Jones,P., Zhu,W., Julian,R.K.Jr and Apweiler,R. (2005) Second proteomics standards initiative spring workshop. *Expert Rev. Proteomics*, **2**, 287–289.
8. Omenn,G.S. (2004) The Human Proteome Organization Plasma Proteome Project pilot phase: reference specimens, technology platform comparisons, and standardized data submissions and analyses. *Proteomics*, **4**, 1235–1240.
9. Martens,L., Van Damme,P., Van Damme,J., Staes,A., Timmerman,E., Ghesquiere,B., Thomas,G.R., Vandekerckhove,J. and Gevaert,K. (2005) The human platelet proteome mapped by peptide-centric proteomics: a functional protein profile. *Proteomics*, **5**, 3193–3204.
10. Van Damme,P., Martens,L., Van Damme,J., Hugelier,K., Staes,A., Vandekerckhove,J. and Gevaert,K. (2005) Caspase-specific and nonspecific *in vivo* protein processing during Fas-induced apoptosis. *Nat. Methods*, **2**, 771–777.
11. Flikka,K., Martens,L., Vandekerckhove,J., Gevaert,K. and Eidhammer,I. (2005) Improving the reliability and throughput of mass spectrometry based proteomics by spectrum quality filtering. *Proteomics*, in press.
12. Stephan,C., Reidegeld,K., Meyer,H.E. and Hamacher,M. (2005) HUPO Brain Proteome Project Pilot Studies: bioinformatics at work. *Proteomics*, **5**, 2716–2717.
13. Lipscomb,C.E. (2000) Medical Subject Headings (MeSH). *Bull. Med. Libr. Assoc*., **88**, 265–266.
14. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
15. Phan,I.Q., Pilbout,S.F., Fleischmann,W. and Bairoch,A. (2003) NEWT, a new taxonomy portal. *Nucleic Acids Res*., **31**, 3822–3823.
16. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.
17. Garavelli,J.S. (2004) The RESID Database of protein modifications as a resource and annotation tool. *Proteomics*, **4**, 1527–1533.
18. Creasy,D.M. and Cottrell,J.S. (2004) Unimod: protein modifications for mass spectrometry. *Proteomics*, **4**, 1534–1536.
19. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.
20. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455.
21. Craig,R. and Beavis,R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.
22. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.
23. Jones,P., Vinod,N., Down,T., Hackmann,A., Kahari,A., Kretschmann,E., Quinn,A., Wieser,D., Hermjakob,H. and Apweiler,R. (2005) Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics*, **21**, 3198–3199.

2.4.2.3. Which data types to make available through public repositories?

SHORT COMMUNICATION

Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories

Lennart Martens[1], Alexey I. Nesvizhskii[2], Henning Hermjakob[3], Marcin Adamski[4], Gilbert S. Omenn[4], Joël Vandekerckhove[1] and Kris Gevaert[1]

[1] Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium
[2] Institute for Systems Biology, Seattle, WA, USA
[3] EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[4] Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

With the human Plasma Proteome Project (PPP) pilot phase completed, the largest and most ambitious proteomics experiment to date has reached its first milestone. The correspondingly impressive amount of data that came from this pilot project emphasized the need for a centralized dissemination mechanism and led to the development of a detailed, PPP specific data gathering infrastructure at the University of Michigan, Ann Arbor as well as the protein identifications database project at the European Bioinformatics Institute as a general proteomics data repository. One issue that crept up while discussing which data to store for the PPP concerns whether the raw, binary data coming from the mass spectrometers should be stored, or rather the more compact and already significantly processed peak lists. As this debate is not restricted to the PPP but relates to the proteomics community in general, we will attempt to detail the relative merits and caveats associated with centralized storage and dissemination of raw data and/or peak lists, building on the extensive experience gained during the PPP pilot phase. Finally, some suggestions are made for both immediate and future storage of MS data in public repositories.

Received: October 27, 2004
Revised: January 14, 2005
Accepted: March 1, 2005

Keywords:
Bioinformatics / Databases / Mass spectrometry

The completion of the human genome project, with the corresponding rise of the field of proteomics, led to the creation of the HUPO projects as the next major collaborative scientific enterprise in the life sciences [1]. In order to achieve the high-aiming goals of these projects in a reasonable time frame, collaborations between multiple labs around the world have been set up, with each of these labs analyzing standard samples using distinct protocols and hardware. The Plasma Proteome Project (PPP), as the pioneering project in the larger HUPO consortium, is the first of these to have amassed a large body of proteomics data during its recently completed pilot phase [2]. Centralized data storage and subsequent dissemination of these data to the scientific community has been addressed through the initial data collection and management work of Marcin Adamski at the University of Michigan, Ann Arbor [3] and the protein identification database (PRIDE) [4] project of the European Bioinformatics Institute (EBI). During the construction of these resources, a lot of discussion was attributed to the storage of the MS data. In particular the storage of the raw, binary data that the machines report has been discussed thoroughly.

As the question of storing raw data has recently been taken up by editors of proteomics journals as well [5], and furthermore affects the proteomics community at large [6],

Correspondence: Dr. Lennart Martens, Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium
E-mail: lennart.martens@UGent.be
Fax: +32-9264-9484

Abbreviations: EBI, European Bioinformatics Institute; **GNU**, GNU's Not Unix; **GZIP**, GNU ZIP; **PPP**, Plasma Proteome Project; **PRIDE**, proteomics identifications database; **RAID**, redundant assay of inexpensive disks

we here present a series of advantages and limitations inherent to the publication of raw data compared to processed peak lists, building on the unique experiences obtained through the PPP.

There seems to be a general consensus in the proteomics community today to request submission of the source data on which reported identifications are based [5]. This will allow other researchers to verify and validate the published conclusions independently. Publishing source data also has the benefit of allowing additional (computational) analyses by other researchers, which could lead to the uncovering of new, biologically relevant information that was missed in the original analysis.

These source data can take a number of forms, but by far the most common representations are either the proprietary, binary "raw" formats that the mass spectrometers churn out during their analyses or the text-based, processed peak lists that are typically submitted to search engines for identification of the peptides that produced those spectra. In the case of fragmentation spectra, the peak lists contain the parent peptide m/z and charge (if the charge is known) and a listing of measured m/z values and their intensities for the fragment peaks. Search engines then attempt to match these fragment peaks to *in silico* generated fragmentation spectra of all peptides in a search database. The peak lists are often called MS/MS spectra and due to the extensive automation of acquisition software, they are often the only format encountered by researchers. These files can take a variety of formats, yet all are essentially text-based, small (a few kilobytes *per* file), readily readable by both humans and software programs and easily compressible (two-fold to three-fold compression ratios are routine using GNU ZIP (GZIP) (GNU – GNU's Not Unix)). Additionally, each of these peak list formats can conveniently be transcribed in any other format. A few common examples are SEQUEST files (dta), Micromass peak lists (pkl), and MASCOT Generic Format files (mgf). There is a slight variability in the amount of information these different formats can accommodate, but in general conversion between formats tends to be conservative. Furthermore, the mzData format, a community standard recently developed by the HUPO Proteomics Standards Initiative (PSI) [7] that elicits broad support among both instrument and software vendors, will ultimately eliminate the need for these format conversions.

As noted above, peak lists present an already processed view on the originally recorded data. Typically proprietary, vendor-supplied software is used to extract these peak lists from the raw data. Frequently applied processing techniques during this extraction phase include noise-filtering, centroiding, deconvolution, and deisotoping of the peaks. As there is no standard protocol for these processing steps, problems often arise because what one scientist regards as standard processing might seem "lossy" conversion to another, leading some to label these peak lists as an unfit distribution medium for MS data.

The raw data formats in contrast are much larger in size (typically well above 10 MB *per* file) and are usually stored in a proprietary, binary format. This makes the files impractical to read for both users and third-party software programs, all the more so because the exact format description is typically not disclosed by the vendors. Since the binary format can already be a compressed representation of the data, standard compression algorithms such as GZIP do not always reduce the size of these files. A simple analysis was performed to illustrate both size differences and the effects of data compression (Fig. 1). The much larger size of the raw data does, however, allow these files to contain much more information than peak lists. Raw files contain all the individual peaks as registered by the instrument detector and, for LC-MS machines, can store elution profiles and times for the LC part. Depending on the vendor and make of the machine, other useful instrument-related information can be stored in these files as well.

Recently, several interesting developments have been described that can put this wealth of additional information in raw files to good use [8, 9]. The key to interpreting these raw data directly has been the development of specific software to parse the binary content of these raw files into intelligible data, a tedious and time-consuming task that typically needs to be redone each time a new machine or a new version of an existing machine or its operating software appears. Furthermore, this reverse-engineering of a proprietary format is typically frowned upon by vendors. Next to the above-mentioned caveats associated with proprietary raw data formats, there is also the very real problem of "aging" that comes with any binary formatted data. As time goes by, support for certain formats tends to evaporate and within the space of several years, readers can no longer be found for the format. A detailed review of the issues concerning proprietary data formats and science can be found in [10].

The mzXML format of the Institute of Systems Biology [11], designed as an intermediate format between raw data and peak lists, could bring some solace if it were supported by vendors, but a more pervasive effort on behalf of the entire community to standardize raw data formats is more likely to succeed in eliciting such global support.

When it comes to storing mass spectrometric data in proteomics data repositories, the discussion tends to focus on an "either-or" decision. Most proponents for the storage of raw data currently have (limited) facilities to parse this kind of data, and are therefore able to exploit the richer information therein. The other camp, which advocates the storage of the processed peak lists, tends to lack this software, making the raw data essentially inaccessible to them (unless they happen to possess the particular, proprietary instrument software that allows the transformation to peak lists). It is our opinion that the choice should not be an exclusive one. In fact, we are convinced that both formats have a distinct and additive value at this time and as such fulfill complementary roles.

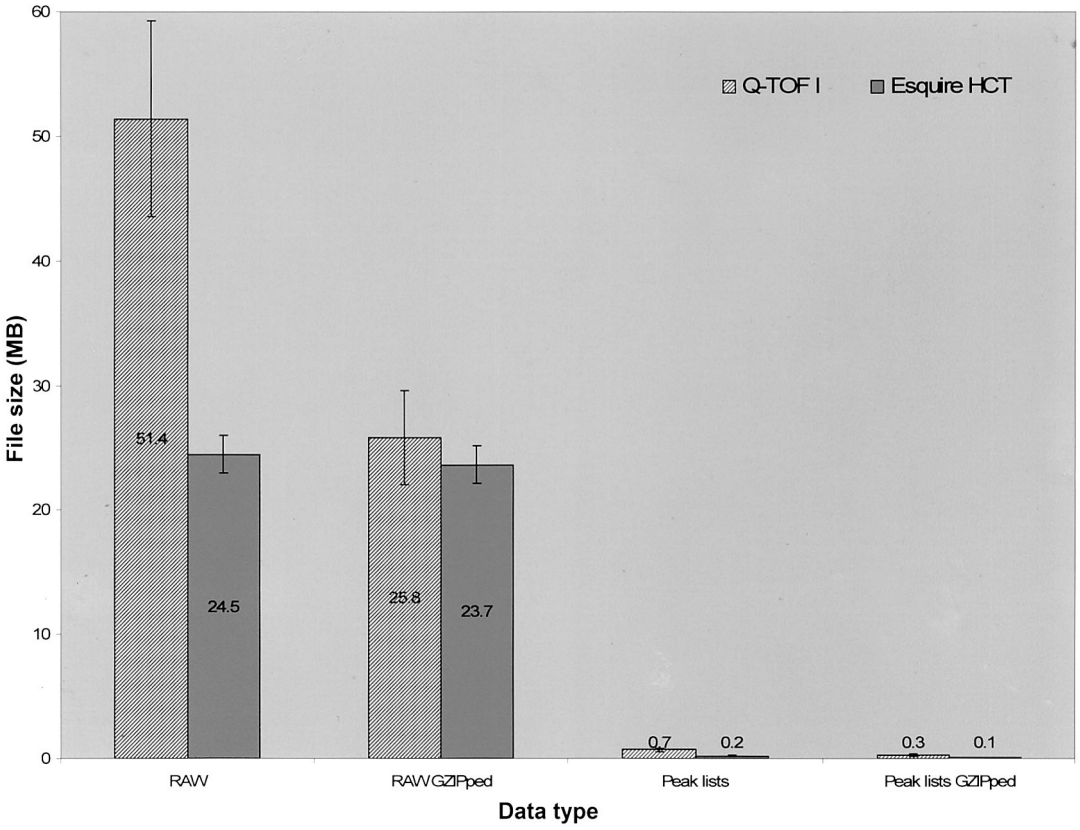


Figure 1. Comparing compressed and uncompressed file sizes for RAW data and the corresponding peak lists. Figures for the data are based on the averages of multiple separate files for each measurement. Error bars denote one SD on the averages. For the raw data, the sizes were averaged over ten individual files. Q-TOF I (Micromass, Cheshire, UK) peak list data consist of 720 individual files, the Esquire HCT (Bruker Daltonik, Bremen, Germany) IT peak list data count 1050 distinct files. Both file sets were grouped into ten subgroups, with each subgroup corresponding to the spectra extracted from a single parent raw file. File format chosen for the peak lists was the intermediately verbose MASCOT Generic Format (http://www.matrixscience.com/help/data_file_help.html). Peak lists have been tarred by GNU tar (http://www.gnu.org) to compensate for size-bloating due to the minimal file size limit of the NTFS file system. Compression for both RAW files and peak lists was done using GZIP with default compression settings. Note the extreme difference in file sizes between raw data files and peak lists. Also notable is the difference in compression efficiency between Q-TOF I RAW files and their Esquire HCT counterparts, especially since the compressed results are highly similar, indicative of a built-in compression in the Esquire HCT files. Compressibility of the peak lists can be deduced from the data labels and is always greater than 50%.

When a reevaluation of the peak lists using a different search algorithm or using a newer sequence database as search base is the scope of the research done with the original data, peak lists typically are the most readily accessible and efficient sources of MS data. For more advanced purposes however, such as obtaining large training sets for machine learning approaches for the prediction of peptide elution times [12] or, in the case of quantitative proteomics experiments based on stable isotope labeling [8], the raw formats present the only data source rich enough for these analyses.

Therefore, in the PPP, peak lists are part of the core data structure, whereas submission of raw files is considered an optional yet highly encouraged addition. The reason for this optional inclusion of raw data is purely technical in origin, as the sheer size of the files involved pushes infrastructure requirements for both storage of the data and their subsequent distribution to their limits.

Typically, funding for these infrastructure issues is evaluated using a standard cost/benefit model, yet for raw data files, the costs will surely outweigh the benefits in the short term. Storing raw files will require large amounts of disk space, which typically should be made redundant (*e.g.*, using RAID systems), thus disk space requirements will be at least twice the size of the data. Back-ups of this amount of data also present a nontrivial challenge. Due to typical low compression ratios, the amount of uncompressed tape media space (which tends to be more expensive than hard

drive space) required will be roughly equivalent to the total data size. The distribution of the data after they have been successfully stored, also accounts for a large part of the cost involved since bandwidth does not come free, either. As an illustration of the data storage requirements, we consider the raw data for a single ICAT [13] or COFRADIC [14] run through a complete proteome (30–40 separate LC-MS/MS runs, with a 2 h gradient each) to have a compressed size of roughly 1.5 GB for older or less sophisticated machines, up to a massive 45 GB for newer, state-of-the-art instruments! It can be expected that future machines will generate even larger files as instrument accuracy and resolution increases. Put in perspective, a single proteome thus requires at least three times as much storage space as the NCBI nonredundant protein database (ftp://ftp.ncbi.nih.gov/blast/db/nr.tar.gz) in FASTA format, or three times as much as the full Swiss-Prot database [15] in the native text format! And although a 100 GB low end hard disk can currently be purchased for about US $100, a conservative cost estimate from the EBI averages to a total cost of US $2000 *per* 100 GB stored for data on a public high-availability FTP server, including distribution and back-up costs!

Even though a truly distributed system (every lab hosting its own raw data) maximizes cost-efficiency through distribution of both the storage and bandwidth cost, it is typically undesirable in the long run as the turn-over for availability of academic sites tends to be quite high. The installation of centralized repositories, located at dedicated institutes such as the EBI or the National Center for Biotechnology Information (NCBI), would be far more reliable in the long run, yet these organizations typically suffer from a lack of resources to host this amount of data. Compared to sequence databases, for instance, the growth in data storage requirements (and hence the rise of the cost) will be far greater for raw data, whereas the benefits (typically calculated in number of downloads or resulting publications) will most probably be less. The lack of open formats for the raw data adds to the difficulty of establishing funding for centralized repositories, which brings us to a catch-22: for a true incentive towards routine dissemination of raw data for published papers, we need open standards for the data formats used, but in order to push such open standards on the vendors, a large user community is needed that can actively define these standards as well as demand support for them from the vendors.

As a conclusion, the following recommendations can be made concerning the dissemination of MS data: (1) peak lists should be made available by default. There is no reason not to make these publicly available, and there are no real storage or distribution issues to be considered. (2) raw data have some clear benefits over peak lists, yet currently lack both standardized formats as well as the required infrastructure for centralized storage and distribution. Therefore, information on how to obtain raw data should at the very least be referenced in the published results for the time being. This can easily be done by providing links to individual lab websites from the journal websites (note that this is a version of the "truly distributed system" discussed above). (3) Efforts should be started at centralized repositories to create the necessary infrastructure so that in the mid- to long-term, source data will preferentially be submitted in the raw format. Meanwhile, (4) vendor support should be enlisted for open formats or at least open access to software tools that allow users to read and interpret the different formats of raw data. Since these latter developments are mutually dependent, the most important breakthrough to achieve seems to be the establishment of centralized repositories. Perhaps some lessons can be learned in this respect from the microarray community, as they have faced (and largely overcome) similar problems in the recent past [16].

The authors would like to thank David States for sharing his experiences from the data gathering efforts executed by his core bioinformatics unit at the University of Michigan, Ann Arbor; Ilan Beer for interesting discussions and his views on the subject matter; Peter Stoehr for the storage cost estimates; Jimmy Eng for contributing raw data file sizes for several instruments; and An Staes for her help in preparing the chart in Fig. 1. K.G. is a Postdoctoral Fellow and L.M. a Research Assistant of the Fund for Scientific Research-Flanders (Belgium) (F.W.O. Vlaanderen). Parts of the data used in this paper were generated in the context of the GBOU-research initiative (Project number 20204) of the Flanders Institute of Science and Technology (IWT).

References

[1] Hanash, S., Celis, J. E., *Mol. Cell. Proteomics* 2002, *1*, 413–414.

[2] Omenn, G. S., *Proteomics* 2004, *4*, 1235–1240.

[3] Adamski, M., Blackwell, T. W., Menon, R., Martens, L., Hermjakob, H., Taylor, C. F., Omenn, G., States, D., *Proteomics* 2005, *5*, this issue.

[4] Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C. F., States, D., Gevaert, K., Vandekerckhove, J., Apweiler, R., *Proteomics* 2005, *5*, this issue.

[5] Carr, S. A., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., Nesvizhskii, A., *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[6] Prince, J. T., Carlson, M. W., Wang, R., Lu, P., Marcotte, E. M., *Nat. Biotechnol.* 2004, *4*, 471–472.

[7] Orchard, S., Hermjakob, H., Randall, K. J., Jr., Runte, K., Sherman, D., Wojcik, J., Zhu, W., Apweiler, R., *Proteomics* 2004, *4*, 490–491.

[8] Li, X. J., Zhang, H., Ranish, J. A., Aebersold, R., *Anal. Chem.* 2003, *75*, 6648–6657.

[9] Beer, I., Barnea, E., Ziv, T., Admon, A., *Proteomics* 2004, *4*, 950–960.

[10] Wiley, H. S., Michaels, G. S., *Nat. Biotechnol.* 2004, *22*, 1037–1038.

[11] Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B. *et al.*, *Nat. Biotechnol.* 2004, *22*, 1459–1466.

[12] Petritis, K., Kangas, L. J., Ferguson, P. L., Anderson, G. A., Pasa-Tolic, L., Lipton, M. S., Auberry, K. J. *et al.*, *Anal. Chem.* 2003, *75*, 1039–1048.

[13] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., Aebersold, R., *Nat. Biotechnol.* 1999, *17*, 994–999.

[14] Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G. R., Vanderkerckhove, J., *Nat. Biotechnol.* 2003, *21*, 566–569.

[15] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E. *et al.*, *Nucleic Acids Res.* 2004, *32 Database issue*, D115–D119.

[16] Ball, C. A., Sherlock, G., Brazma, A., *Nat. Biotechnol.* 2004, *22*, 1179–1183.

2.5. Application to human platelet proteomics

2.5.1. An all-round showcase

The meta-analysis of four distinct experiments comprising three different COFRADIC views of the human platelet proteome is an interesting showcase for the results discussed above. Two N-terminal COFRADIC experiments, one cysteine COFRADIC experiment and one methionine COFRADIC experiment were combined and re-searched to yield as complete a platelet proteome as was possible. It also allowed the study of potential complementarity or overlap in the different COFRADIC procedures.

In order to manage the data from these projects and later re-search and integrate them ms_lims was instrumental. It also made analyses within and across COFRADIC techniques easy and fast. DBToolkit-derived databases for N-terminal COFRADIC increased the total number of identified proteins for the two corresponding projects by 50%, with many of these additional identifications providing precise information about protein processing. The datasets (spectra and identifications) were also the first ones ever to be made publicly available via the PRIDE system and have correspondingly been stored as PRIDE accession numbers 1, 2 and 3.

A later analysis of the complete N-terminal dataset with the spectrum classifier tool revealed a list of unidentified yet high quality spectra that were subjected to a detailed analysis for unexpected modifications. This resulted in the detection of the cyclization of N-terminal S-carbamoylmethylcysteine as an important yet unexpected artefactual modification. This discovery has since been taken into consideration within each N-terminal COFRADIC project (see section 2.3 above).

Taken together, the application of the tools discussed in the previous sections to this dataset allowed the identification and publication of the single largest set of identified proteins in human blood platelets to date.

2.5.2. Publication

REGULAR ARTICLE

The human platelet proteome mapped by peptide-centric proteomics: A functional protein profile

Lennart Martens, Petra Van Damme, Jozef Van Damme, An Staes, Evy Timmerman, Bart Ghesquière, Grégoire R. Thomas, Joël Vandekerckhove, Kris Gevaert

Department of Medical Protein Research, Flanders Interuniversity Institute for Biotechnology, Department of Biochemistry, Ghent University, Ghent, Belgium

Received: July 28, 2004
Revised: November 8, 2004
Accepted: November 15, 2004

Several studies have been published in which holistic approaches were used to characterise the proteome and transcriptome of human platelets. The key intent being that a deeper understanding of the normal and aberrant physiological functions of platelets can only be achieved if most biomolecular building blocks are mapped. Here we present the application of recently developed novel technologies that overcome some of the shortcomings of gel-based proteomics. Central in our approach is the so-called combined fractional diagonal chromatography (COFRADIC)-technology in which sets of representative peptides are sorted in a diagonal RP chromatographic system through a specific modification of their side chain. In this study we combined three different COFRADIC sorting techniques to analyse the proteome of human platelets. Methionyl, cysteinyl and amino terminal peptides were isolated and analysed by MS/MS. Merging the peptide identifications obtained after database searching resulted in a core set of 641 platelet proteins, which comprises the largest set identified today. In comparison to previously published platelet proteomes, we identified 404 novel platelet proteins containing a high number of hydrophobic membrane proteins and hypothetical proteins. Furthermore we discuss the observed characteristics and potential benefits of each of the different COFRADIC technologies for proteome analysis and highlight important issues that need to be considered when searching sequence databases using data obtained in peptide-centric, non-gel proteomics studies.

Keywords:
Bioinformatics / Combined fractional diagonal chromatography / Human platelets / Non-gel proteomics

1 Introduction

Blood platelets are vital for maintaining a closed blood flow through the circulatory system. In normal physiological conditions, platelets respond to a breach in this system by adhering to the lesion site of a vessel wall and by forming a fragile, primary thrombus. This primary thrombus becomes firmer upon fibrin formation by thrombin cleavage of fibrinogen and thickens into a stiff, secondary thrombus. Over the course of wound healing, a fraction of this haemostatic thrombus is degraded by fibrinolysis [1]. Aberrant platelet functions have been associated with various diseases, both acquired and inherited. These diseases, characterised by errors in platelet production (thrombocytopenia and thrombocytosis) and errors in forming, storing and releasing their molecules, cause platelets either to fail to respond to injuries or to incorrectly stimulate clot formation at uninjured sites. Some examples of inherited functional platelet disorders include the Bernard–Soulier syndrome (defects of the glycoprotein Ib-IX-V complex), the von Will-

Correspondence: Professor Kris Gevaert, Department of Medical Protein Research, Flanders Interuniversity Institute for Biotechnology, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium
E-mail: kris.gevaert@UGent.be
Fax: +32-92-649-496

Abbreviation: COFRADIC, combined fractional diagonal chromatography

enbrand disease (disorder of vWF) and the Wiskott–Aldrich syndrome (decreased levels or absence of WASP) [2]. In order to fully understand the ongoing biochemical processes in platelets, large-scale proteomics [3–6] and transcriptomics [7] studies have been performed. The general idea being that a large catalogue of platelet proteins could form the basis for further experiments, thus accentuating the explorative role of these studies. While in one of the largest ongoing 2-D gel based platelet proteome studies today about 2300 distinct protein spots have been visualised [4, 6], thus far only 411 different platelet proteins could be identified after analysing the observed protein spots by LC-MS/MS analysis.

Intrinsic shortcomings of 2-D PAGE, important instrumental improvements and the increasing availability of genome sequences have recently led to a new type of proteomics; gel-free or non-gel proteomics. Here, isolated proteomes are first digested (mostly using a highly specific protease) after which either a specific class of peptides is affinity-isolated for MS/MS analysis or the whole, extremely complex peptide mixture is separated by a multidimensional approach prior to analysis (for a recent review see [8]).

We have recently developed a technique for isolating representative peptides out of a tryptic digest of a proteome. It is based upon the principal of diagonal chromatography [9, 10]; between two identical, chromatographic separations a chemical or enzymatic reaction is performed such that specific sets of peptides shift away from their original elution position and are collected. From a proteomic point-of-view such representative peptides preferably contain rare amino acids that are uniformly distributed throughout a proteome. Through this selective isolation technique, the complexity of the final analyte mixture is reduced while the chance that every original protein is represented by at least one peptide remains high, thus achieving minimal complexity while retaining maximal coverage.

Probably the simplest chemical reaction is used for the isolation of methionyl peptides; after a primary separation, the collected peptide fractions are treated with hydrogen peroxide which oxidises methionine to its sulphoxide and thus renders methionyl peptides more hydrophilic. During a secondary, identical separation, methionyl peptides undergo a hydrophilic shift and are collected for MS/MS analysis [11]. Cysteine residues in proteins are first modified using Ellman's reagent, the proteins are subsequently trypsinised and the peptide mixture is fractionated a first time. Prior to the secondary fractionation, the hydrophobic thionitrobenzoic acid group is removed from the cysteine backbone by a reducing agent and cysteinyl peptides undergo a hydrophilic shift, which allows easy isolation from the bulk of non-cysteinyl peptides [12]. In both cases – isolation of methionyl or cysteinyl peptides – the number of separations is reduced by combining multiple primary fractions prior to the secondary separation. Hence, we have called this technology 'combined fractional diagonal chromatography' (COFRADIC).

The highest complexity reduction, however, is obtained by isolating only the amino terminal peptides out of the proteome digest. In this case, the peptides of interest are not shifted but all the internal ones are. After chemically modifying the proteins and subsequent trypsin digestion, only the internal peptides will contain a free α-amino group, which is modified by trinitrobenzenesulphonic acid prior to the secondary separation. The modified internal peptides thus become very hydrophobic (a trinitrophenyl moiety is attached at their terminal amine) and shift out of the primary collection interval. *N*-terminal peptides remain stationary and are collected [13].

For the first time we have applied the three different COFRADIC protocols for the characterisation of one proteome: that of nonstimulated human platelets. After combining the results, a core set of 641 platelet proteins was identified. We here discuss notable differences observed after identifying MS/MS spectra obtained from the three different types of isolated peptides and difficulties encountered when identifying proteins in large databases using only one or a few peptides. A meta-analysis of the core platelet proteome was performed and we noticed that although a large number of proteins were classified as nuclear proteins, only a small part did in fact have oligonucleotide binding properties. Finally, as this proteome contains a high number of transmembrane proteins (87 different proteins or 13.5%) of which the majority was never visualised/identified on a 2-D gel, we here demonstrate that COFRADIC is significantly less biased against identifying hydrophobic proteins compared to 2-D gel based analysis.

2 Materials and methods

2.1 Preparation of a human platelet proteome

A platelet-rich suspension isolated by thrombocytapheresis was obtained from the Red Cross Blood Transfusion Centre Oost-Vlaanderen, Ghent, Belgium. In each proteome analysis 50×10^9 platelets were used. Residual white and red blood cells were first removed by centrifugation of this platelet-rich suspension at $300 \times g$ for 10 min. The isolated supernatant containing the platelets was then centrifuged at $1000 \times g$ for 10 min and the obtained pellet of blood platelets was resuspended in 40 mL washing buffer (0.2 g/L KCl, 0.2 g/L $MgCl_2$, 8 g/L NaCl, 1 g/L D-glucose and 70 mg/L EGTA in 25 mM sodium phosphate at a pH of 7). The presence of EGTA in the buffer system prevented artefactual activation of the platelets. This washing procedure was repeated twice and the platelets were finally resuspended in 10 mL of washing buffer and lysed by adding 10 mL of 0.5% Triton X-100 in 25 mM of sodium phosphate buffer (pH 7.5) containing a cocktail of protease inhibitors (Roche Diagnostics, Mannheim, Germany). Lysed platelets were put on ice for 30 min after which cellular debris was removed by centrifugation (10 min at

10 000 × *g*). For each COFRADIC experiment, 500 μL of this protein mixture (corresponding to 1.25 × 10^9 platelets) was used.

2.2 COFRADIC isolation of methionyl, cysteinyl and amino terminal peptides

COFRADIC-based isolation of representative peptides was performed as described elsewhere [11–13].

2.3 Automated LC-MS/MS analysis and identification of the generated MS/MS spectra by MASCOT

Secondary fractions containing sorted methionyl or cysteinyl peptides were pooled prior to LC-MS/MS analysis in order to reduce the number of analyses [11, 12], whereas fractions containing sorted amino terminal peptides were not pooled and each secondary fraction was analysed separately [13]. Dried peptides were redissolved in 20 μL of solvent A (0.1% formic acid and 2% ACN in water; ACN and water were of 'Baker HPLC analysed' quality (Mallinckrodt Baker, Deventer, The Netherlands). Ten microlitres of this peptide mixture were injected on a 0.3 mm id × 5 mm trapping column (PepMap; LC Packings, Amsterdam, The Netherlands) at a flow rate of 20 μL/min (total loading time of 5 min) using a CapLC system (Micromass, Cheshire, UK). By switching the stream valve, the trapping column is back-flushed with a binary solvent gradient, which is started simultaneously with the injection cycle. The sample is thereby loaded onto a nanoscale RP C18 column (0.75 id × 150 mm PepMap™ column; LC Packings). Peptides were eluted from the stationary phase using a linear gradient from 0 to 100% solvent B (70% ACN in 0.1% formic acid) applied over a period of 25 min. The solvent delivery system was set at a constant flow of 5 μL/min and using a 1/25 flow splitter, 200 nL/min of solvent was directed through the nanocolumn.

The outlet of the nanocolumn was in-line connected with a distal metal-coated fused silica PicoTip™ needle (PicoTip™ FS360-20-10-D-C7, New Objective, Woburn, MA, USA), which was placed in front of the inlet of a Q-TOF1 mass spectrometer (Micromass). Automated data-dependent acquisition was initiated 15 min after the stream valve was switched. The acquisition parameters were such that only doubly and triply charged ions were selected for fragmentation and fragmentation spectra were merged over a period of 8 s. The stream valve was switched back 51 min after the start of the injection cycle.

The acquired CID-spectra were automatically converted to a MASCOT [14] acceptable format using the ProteinLynx program of Micromass' MassLynx software (version 3.4). Different sequence databases were used in sequence to identify the isolated peptides. In all three COFRADIC experiments, MASCOT first searched in the regular IPI human protein database, version 2.31 (ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/ipi.HUMAN.dat.gz). For the methionyl and cysteinyl peptides, the IPI human database was also searched in MASCOT 'No Enzyme' mode. In this mode, all residues are considered potential cleavage sites and thus all possible peptide sequences can be considered for spectrum matching. In the case of amino terminal COFRADIC two additional in-house developed databases were searched: first an amino terminal 'ragged' version of the truncated IPI human database where the truncation length for each IPI entry was set to 100 amino terminal amino acids. This database allowed the identification of post-translational protein processing such as signal peptide cleavage. The second database was also an amino terminal ragged version of the IPI human database, yet this time the ragging occurred on the full protein sequences. The truncation and ragging was performed using the DBToolkit software (http://www.proteomics.be/bioinfo/lm/dbtoolkit) as described previously [13]. To avoid the possibility that one MS/MS spectrum was linked to identical sequences stored in different databases, the follow-up search option that is available in the MASCOT daemon tool was used in such a way that only the MS/MS spectra that were not identified at the 95% confidence level in one database, were used to search in the larger databases. The search order employed for the methionyl and cysteinyl peptides was first regular IPI human, then IPI human with the 'No Enzyme' setting. For the amino terminal COFRADIC the search order used was first the regular IPI human, then the ragged IPI human database truncated to 100 amino terminal residues and finally the ragged IPI human database without truncation.

Depending upon the type of peptide analysed, the MASCOT search parameters were set differently as previously described [11–13]. For all searches both the precursor mass tolerance and the fragment mass tolerance were set to 0.3 Da. Instrument setting was always 'ESI Q-TOF'. For the regular IPI database searches, as well as the 'No Enzyme' searches, the number of allowed missed cleavages was set to 1. For the truncated databases, no missed cleavages were allowed for the searches, since one missed cleavage was already allowed during construction of the truncated peptide databases. For the *N*-terminal COFRADIC spectra, the fixed modifications were set to acetylated lysines and carbamidomethyl cysteines. The allowed variable modifications were acetylated *N*-termini, deamidation for asparagines and glutamines, oxidation of methionines and pyroglutamate for *N*-terminal glutamines. For the methionine and cysteine COFRADIC there were no fixed modifications and the following variable modifications: deamidation for asparagines and glutamines, oxidation of methionines and pyroglutamate for *N*-terminal glutamines.

The DAT result files of MASCOT were automatically queried using in-house developed software tools and only MS/MS spectra that were identified by a score that exceeded the identity threshold score of MASCOT at the 95% confidence level were retained. The retained spectra were subsequently manually validated and only spectra that held a high number of typical fragment ions were considered as positively identified (typically about 50% of b and y ions were

present). The identified peptides were automatically stored in a MySQL relational database in which links were made to their MS/MS spectra and precursor proteins.

3 Results and discussion

3.1 MS/MS analysis of peptides sorted by different COFRADIC techniques

Since we pursued a comprehensive analysis of the human platelet proteome, we used three different COFRADIC technologies: sorting methionyl peptides [11], cysteinyl peptides [12] and amino terminal peptides [13] respectively. These peptides were analysed by automated LC-MS/MS on a Q-TOF1 mass spectrometer and the obtained peak lists were used as input for MASCOT [14]. Searches were done in version 2.31 of the nonredundant International Protein Index (IPI) database [15], with restriction to human proteins.

The spectra from the *N*-terminal COFRADIC analysis of platelets as published previously [13], as well as those resulting from the cysteine COFRADIC analysis of platelets [12], were re-searched in the IPI database along with the unpublished methionine data. The reason for this re-searching was two-fold: in this way a common search base is employed and, at the same time, this search base is brought up-to-date. IPI was chosen because it provides a minimally redundant view on a number of popular sequence databases (*e.g.,* Swiss-Prot and NCBI nr). The identified peptides/proteins are stored in the PRIDE database and accessible *via* http://www.ebi.ac.uk/pride/. PRIDE experiment accession number 1 points to the identified methionyl peptides, number 2 to the cysteinyl peptides and numer 3 to the *N*-terminal peptides.

We noted a significant difference in the identification efficiency of the obtained spectra (Table 1). Spectra from methionyl peptides clearly have a higher identification efficiency compared to those from sorted cysteinyl and amino terminal peptides. In COFRADIC, methionyl peptides are converted to their hydrophilic sulphoxide derivatives. When analysed by mass spectrometers, such peptides are easily recognised by the readily occurring neutral loss of methanesulphenic acid (CH_3SOH) [16]. However, we also noted that such peptides are impaired for fragmentation of their backbone and generally tend to give fewer and less intense fragment ions. This phenomenon is especially prominent in MALDI PSD [17] (results not shown). In this view, one would expect that MS/MS spectra of such peptides are less informative and thus more difficult to link to peptides. Here, this is clearly not the case. In fact the identification efficiency of spectra from methionyl peptides is about twice as high as that of cysteinyl and amino terminal peptides. Using the same experimental setup, the average identification efficiency of a typical human protein sample digested with trypsin is 30%, which is comparable to the identification efficiency observed for methionyl peptides.

Apparently this hints to the fact that MS/MS spectra obtained from COFRADIC sorted cysteinyl and amino terminal peptides are more difficult to link to peptide sequences by MASCOT compared to those from methionyl peptides.

Possible causes for the apparent less efficient identification of cysteinyl peptides might be due to the chemical reactions employed for their isolation. For instance, the hydrophobic group could have reduced the solubility of the Cys-modified peptides resulting in a loss of such peptides during the primary separation. Between the primary and secondary run, the hydrophobic thionitrobenzoic acid group is removed from the cysteine backbone by a reduction reaction using phosphines [12]. If such reaction mixtures are kept at room temperature for a long time, the phosphines become oxidised and do not protect the free thiol groups anymore. Therefore, disulphide bridges might spontaneously occur [18] leading to homo- or hetero-dimeric peptides that might not get isolated for further LC-MS/MS analysis because of their aberrant retention time or when analysed by MS/MS, their sequence(s) will not be retrieved in databases since their parent mass reflects the combined masses of the peptides and as such is too high.

Sorted amino terminal peptides are not archetypal tryptic peptides: they all end on an arginine residue, their amines are acetylated and free thiol groups are alkylated [13]. We performed a statistical study on the fragmentation behaviour of these peptides compared to 'normal' tryptic peptides and could not really find a dissimilarity indicating that the chemical nature of these peptides does not modify CID fragmentation such that peptide identification is hampered (data not shown). A more likely cause for the reduced identification efficiency is due to the fact that the mass spectrometer encounters too much 'dead time' when analysing amino terminal peptides. During COFRADIC, methionyl and cysteinyl peptides are induced to shift out of their original collection interval by making them more hydrophilic. An interesting, secondary effect evoked here is the fact that sort-

Table 1. Comparison of the identification efficiencies and the number of identified platelet proteins using the three types of COFRADIC sorting strategies

	% of identified spectra	Number of identified proteins(−isoforms)	Number of identified proteins(+isoforms)	Average number of spectra/protein	Average number of peptides/protein
Methionyl peptides	35.9	375	768	3.41	1.88
Cysteinyl peptides	20.3	157	349	3.45	1.99
N-terminal peptides	15.9	345	592	5.68	1.54

ed peptides elute over a significantly larger interval than during the primary run (typically about five times larger). In order to reduce the number of LC-MS/MS runs, the secondary fractions containing sorted methionyl and cysteinyl peptides are combined prior to the final LC-MS/MS analysis. This has the additional effect of spreading the peptides in the LC separation, effectively 'filling' the gradient such that peptides are presented to the mass spectrometer during the entire separation interval. Amino terminal COFRADIC can be regarded as the reverse of the two other procedures; here the peptides of interest remain stationary while the internal ones are induced to shift out of the primary collection interval. The secondary fractions are thus collected in exactly the same time interval as the primary ones. When these fractions are analysed by LC-MS/MS they elute in a significantly smaller time window than the sorted methionyl and cysteinyl peptides. These almost discrete elution windows will be surrounded by periods of dead time during which no peptides elute, prompting the mass spectrometer to analyse contaminating ions instead. Evidence for this comes from the fact that the average TIC of MS/MS spectra from identified peptides is about three times higher than that of spectra that were not linked to a peptide sequence. Although this could indicate that the unidentified spectra are derived from peptides with very low fragmentation efficiencies, most probably these spectra are from contaminating ions (*e.g.*, for methioninyl peptides this difference in average TIC is only about 1.5). Additionally, the higher sorting efficiency of amino terminal COFRADIC results in a less diverse peptide mixture in each elution window, which causes the mass spectrometer to reanalyse the same peptide ions many times. These effects lead to a larger fraction of unidentifiable spectra, reducing the perceived identification efficiency, and more redundant peptide identifications. Supporting evidence for the latter comes from the fact that although the smallest set of proteins identified by the methionine and the amino terminal COFRADIC approach is highly similar (376 *vs.* 345, Table 1), the average number of spectra identified *per* protein almost doubles in amino terminal COFRADIC whereas the number of identified peptides is lower.

3.2 Different proteins are identified by different COFRADIC sorting procedures

We used MASCOT to identify the peptides and set the identity threshold at the 95% significance level. We verified MASCOT's results and whenever a small number of fragment ions or many rare fragments were identified (typically a, c, x and z ions, nomenclature according to [19]) the identifications were discarded. We finally combined the peptide identifications obtained by following the three different COFRADIC routes and linked them to 673 different protein entries stored in the IPI database; however, in many cases the identified peptides could not be distinguished between different protein entries (mainly isoforms or splice variants).

About two-thirds of all proteins are identified in distinct COFRADIC experiments and the overlap between the different sets of identified proteins is small; for instance less than 6% of all proteins are identified in all three experiments (Fig. 1). We think that this is due to undersampling of ions by mass spectrometers. In a typical COFRADIC experiment the flux of peptides to the mass spectrometer is very high; on average every second a few peptides elute from the RP column and are ionised. In our instrumental setup a peptide typically elutes in a time frame of 30 s and the duty cycle of the mass spectrometer (MS scan followed by MS/MS analysis) is 8 s. This indicates that of the high number of peptides eluting in a 30 s window, only three to four will be fragmented and may lead to peptide identification. Clearly a lot of information is lost this way and one way to overcome this is to reanalyse the sample using *m/z* exclusion lists or to separate the analytes to a higher degree using multidimensional chromatographic steps [20] either before or after the COFRADIC sorting step. This should eventually lead to a higher number of peptides (and proteins) identified and thus to a bigger overlap between the results of the separate COFRADIC approaches.

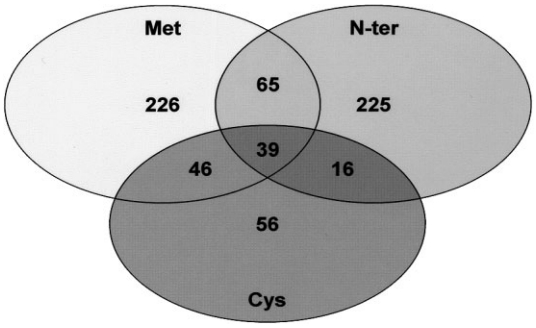


Figure 1. Scheme showing the overlap between the proteins identified by the three different COFRADIC approaches (the number of proteins is indicated).

3.3 Meta-analysis of 641 platelet proteins identified by COFRADIC

Peptides were identified in the IPI database, which is a non-redundant amalgam of databases on the protein level and is expected to become the most resourceful database for proteome studies in the near future [15]. Most non-gel proteomics approaches are peptide-centric; *i.e.*, only small pieces of proteins (typically between 10 and 20 amino acids long) are analysed and linked to the parent protein sequence. Especially when analysing proteomes of higher eukaryotes, such identified peptides can reside in multiple database entries and it becomes very difficult to pick the protein form that was actually present in the original sample. In some cases meta-information can be used (*e.g.*, the cell type or tissue analysed) to narrow down the possibilities but it generally remains difficult to exactly identify a particular database entry using only one or a few peptide sequences. Next to this peptide redundancy issue we found that many proteins

which were termed 'hypothetical' or 'similar to' in their descriptions perfectly matched to well-characterised proteins following homology searches (http://www.ebi.ac.uk/blast2/). Most of these protein entries appeared to be shortened versions (more than 5% in total length) of other, well-characterised proteins.

After cleaning up the original list of 673 protein entries, we end up with a set of 641 different protein entries (Supplementary Table 1). We call this set the core platelet proteome as it might contain many more proteins. Indeed, only in a minor number of cases can we use the identified peptides to distinguish between different isoforms or splice variants (mainly nonhighlighted entries in Supplementary Table 1). Since protein isoforms tend to differ mainly at their extremities we found that *N*-terminal peptides are more efficient at distinguishing between isoforms than methionyl or cysteinyl peptides.

A total of 3771 MS/MS spectra corresponding to 1625 different peptides identified the 641 proteins (Supplementary Table 1). In this study, 401 proteins (62.55%) were identified by a single peptide and of these 307 (47.89% of all proteins) were identified by a single MS/MS spectrum. Recently, there has been some debate in the literature concerning the soundness of such 'one-hit wonders' [21] and in a recent study, Western blots validated the peptide ratios originally predicted by such unique peptides observed in the original non-gel proteomics study [22]. This suggests that platelet proteins that are here identified by a single spectrum or a single peptide cannot simply be discarded and therefore must be seen as part of the proteome. Certainly, if 'interesting proteins' (*e.g.*, potential drug targets or biomarkers) are present as one-hit wonders, the protein's identity can be further verified by 'reverse proteomics'. For instance, one or more methionine-containing tryptic peptides predicted from the protein sequence could be synthesised in their heavy (D, ^{13}C, ^{15}N, ^{18}O) form and spiked in the peptide mixture (*e.g.*, [23]). Since their chromatographic behaviour in the COFRADIC sorting steps can be perfectly predicted, one can target from the synthetic peptides the corresponding light peptides in the complex mixture.

We have compared our non-gel platelet proteome with five lists of platelet proteomes that have been published before [3–6, 24] and together comprise the largest data set obtained on the human platelet proteome thus far. About 62.5% (401 proteins) of the non-gel proteome has not yet been identified in platelets before, which clearly illustrates the discovery power of our non-gel technologies. Interestingly, following comparison of the largest 2-D gel based platelet proteome data set obtained thus far with our non-gel proteome we noticed that only 164 of the proteins identified on gel (out of a total of 411 [4, 6]) were identified in our study. This apparent low overlap is probably due to the aforementioned undersampling of peptide ions by mass spectrometers; peptide ions of the missing proteins were probably present in the analyte mixtures but were missed by the mass spectrometer because of its rather long duty cycle.

One way of characterising a large set of proteins is to classify it according to cellular location and molecular function using the data deposited in the Gene Ontology database (http://www.geneontology.org) [25]. When classified by cellular location, about 64% of the proteins are present in the cytoskeleton, ER, mitochondrion, cytosol and Golgi apparatus while 16% are located in or at the plasma membrane (see below) and 20% in the nucleus (Fig. 2A). The latter figure is intriguing since platelets are devoid of nuclei while in our analysis nuclear proteins appear to form the largest portion of the platelet proteome (Fig. 2A). However, as they are fragments of large progenitor cells (megakaryocytes), it could be that platelets or a fraction of them still hold putative nuclear proteins, particularly those that can shuttle between the nucleus and the cytoplasm. Nevertheless, such nuclear proteins might be the basis for future studies to find out their role in anucleated platelets. Contamination of our platelet preparation by nucleated blood cells is unlikely, as we have not detected histones or histone fragments, known as ubiquitous major nuclear components. On the other hand, the identification of α and β globin (11 and 3 spectra respectively) and β-spectrin (1 spectrum) could point to contamination of our platelet sample by red blood cells. However, the fact that these abundant red blood cell proteins were identified using a very small number of all obtained MS/MS spectra (0.4%) indicates that these proteins were only very lowly abundant in the analysed proteome [26, 27]. Furthermore, the clear absence of other abundant erythrocyte proteins such as alpha-spectrin, ankyrin, proteins bands 3, 4.1 and 4.2 [28] suggests that this contamination is very limited; which means that the very large majority of the proteins mentioned in the core proteome are true platelet components.

After classifying the proteome according to the molecular function of its protein components, enzymes, enzyme regulators, proteins involved in transport, cytoskeletal and structural proteins, signal transducers and chaperones make up the largest part of the proteome (see Fig. 2B). As expected, regulation of gene expression such as transcription and translation appears to be performed by a smaller set of proteins.

3.4 Compositional comparison of the platelet proteome to two proteomes of nucleated cells analysed by COFRADIC

When the platelet proteome is compared to proteomes of actively proliferating cells such as human Jurkat cells (829 proteins identified following COFRADIC analysis, unpublished data) and human neuroblastoma SH-SY5Y cells (656 proteins identified following COFRADIC analysis, unpublished data), we notice a significant difference in the protein localisation and protein function profiles for these human cell types (Fig. 3). While the relative number of proteins present in most subcellular structures (ER, mitochondria, cytosol, Golgi and plasma membrane) is highly similar among the three proteomes, nuclear proteins

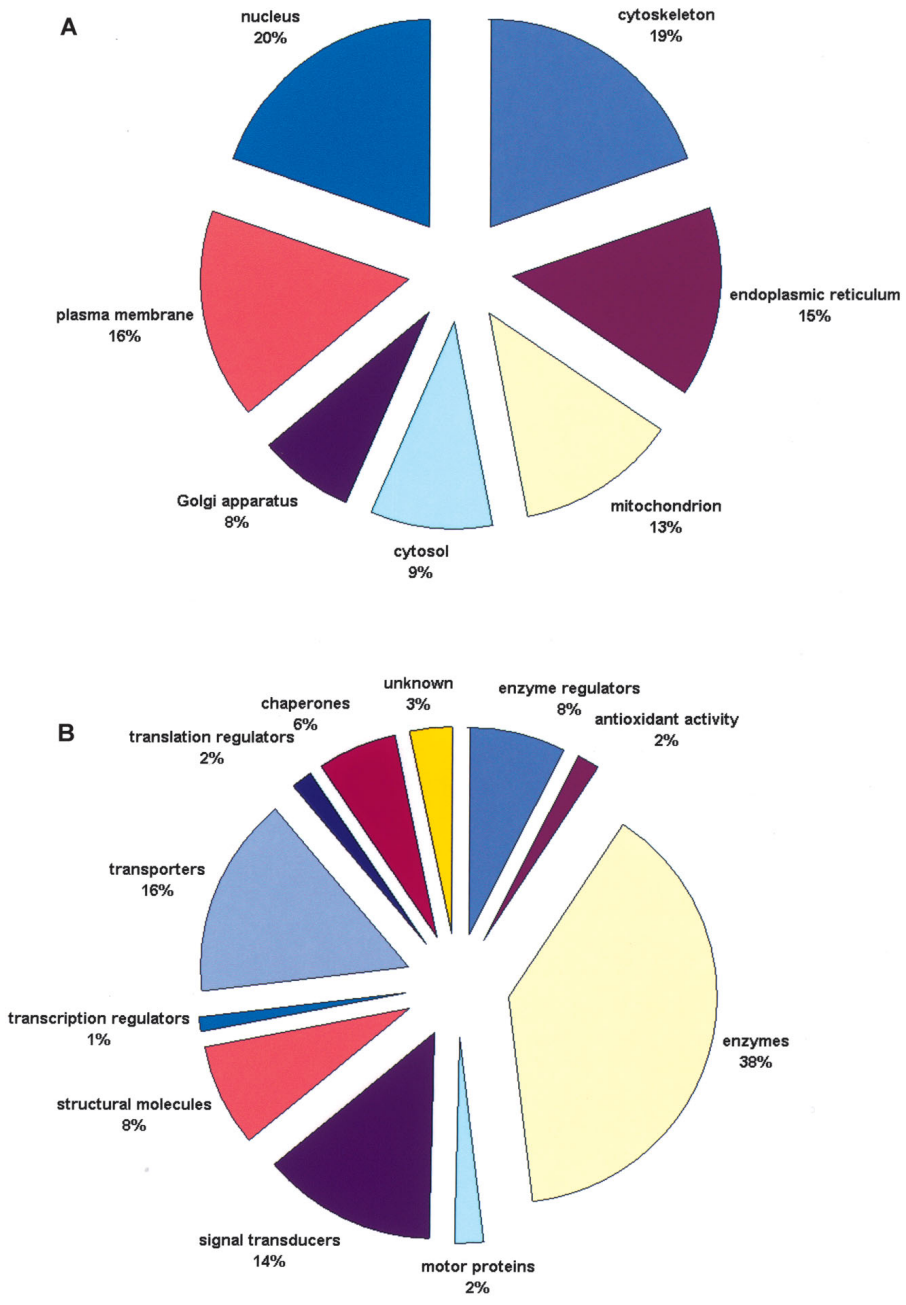


Figure 2. Meta-analysis of the identified platelet proteins. Identified platelet proteins were classified according to their cellular location (pie chart A) and according to their molecular function (pie chart B) using the data stored in the Gene Ontology database (see text). Relative quantity of proteins present in a certain class is indicated. Clearly, cytoplasmic proteins make up the largest part of identified proteins, although 20% of all proteins identified are nuclear (A) while the major part of the identified proteins are enzymes and their regulators (B).

in particular make up a bigger part of the proteomes of nucleated cells compared to that of platelets (Fig. 3A). On the other hand, the number of proteins involved in organising and modulating the cytoskeleton is almost doubled in the proteome of human platelets compared to the two other proteomes (Fig. 3A). This demonstrates an important physiological role of platelets in thrombus formation and consecutive contraction. Activated platelets change their morphology promptly; a process evoked by signalling cascades that eventually leads to a redistribution of cytoskeletal and cytoskeletal-associated proteins and a reconstruction of the actin cytoskeleton. Clearly, a high number of structural and cytoskeletal proteins must be required for this process.

Following a classification of the proteome building blocks according to their known or predicted molecular function it is clear that, as expected, apart from proteins involved in regulating gene transcription, all functional classes are present in highly equivalent numbers in the three proteomes (Fig. 3B).

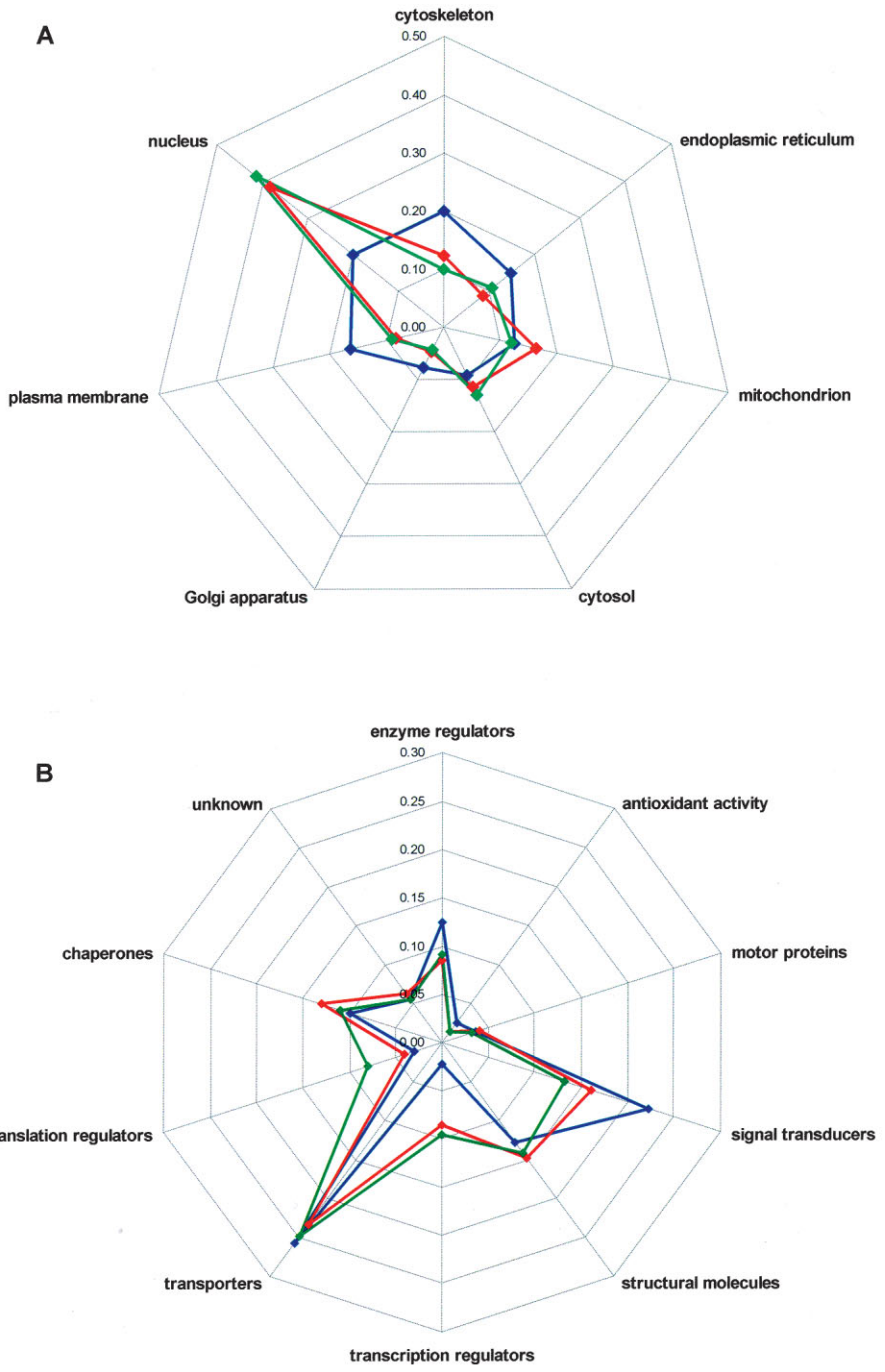


Figure 3. Radar plots comparing the composition of the human platelet proteome to proteomes of nucleated human cells. In both radar plots, the human platelet proteome is indicated by a blue line, whereas the proteomes obtained following gel-free analysis of human Jurkat T-lymphocytes and human neuroblastoma SH-SY5Y cells are indicated by green and red line respectively (unpublished data). Radar plot A shows the relative distribution of the three different proteomes according to the cellular location of their components as indicated in the Gene Ontology database. Relative distribution of the different molecular functions of the identified proteins is indicated in radar plot B (for clarity reasons enzymes (see Fig. 2B) are not considered for this analysis). Axis of both radar plots is normalised to 1.

One of the weaker points of classical 2-D gel-based proteome analysis is the fact that hydrophobic proteins tend to be underrepresented in 2-D gels. This is because such proteins are very hard to extract out of cellular membranes and tend to precipitate near their p*I* during IEF. Suggested remedies for this shortcoming include the use of chaotropic agents such as thiourea [29] and/or nondetergents such as sulphobetaines [30]. Non-gel proteome analytical techniques tend to be not biased towards the identification of soluble proteins as they first chop them up into peptides which are generally more soluble and readily analysable by mass spectrometers. In this proteome study we identified 87 proteins that contain at least one predicted transmembrane helix (predictions were done with TMHMM Server v. 2.0 at http://www.cbs.dtu.dk/services/TMHMM-2.0 [31]). Following database searching, 69 of these were classified as trans-

membrane proteins while only 12 of them have been characterised before in platelet proteomes [3, 4, 6, 24] (Table 2). After calculating the grand average of hydropathicity (GRAVY) values of these membrane proteins, 24 are classified as hydrophobic proteins (positive GRAVY value [32]). These results support the idea that the COFRADIC technologies are less biased towards hydrophobic membrane proteins compared to gel-based technologies. Clearly this opens up the possibility for discovering novel proteins at the surface of platelets which may eventually be considered as novel drug targets for treating or preventing cardiovascular diseases [33].

Upon stimulation with various agents platelets alter their shape, aggregate and release the content of their granules. These processes are governed by signalling pathways that are controlled by the action of kinases and phosphatases. A thorough investigation of the phosphorylation flux in platelets (*e.g.* [34]) characterises the affected proteins but one can typically only 'guess' at the identity of the enzymes that are responsible (*e.g.*, by studying the motif surrounding the phosphorylated amino acid). Knowing which kinases and phosphatases are expressed in platelets may be of larger interest since inhibitory drugs are currently being developed. Our gel-free proteome contains 19 protein kinases and five protein phosphatases (Table 3). Some of these kinases have never been identified in a large-scale proteomics study in platelets and when added to the repertoire of already known platelet kinases [3–6], they may give deeper insight into the kinome of platelets. Interestingly, other large-scale platelet proteome studies identified a relatively higher number of proteins involved in signalling (*e.g.*, several adapter proteins, small G proteins and their regulators) (*e.g.*, [3, 4, 6]) and our non-gel proteome analysis has clearly missed some of them. A possible reason might be the fact that these proteins do not provide methionyl, cysteinyl or amino terminal peptides that fall into the analysis range of our COFRADIC technique (*e.g.*, too hydrophobic or not easy to ionise or fragment). In this view, combining gel-based and non-gel proteome analyses might be valuable for augmenting the coverage of a given proteome.

Finally, this core platelet proteome contains 51 hypothetical proteins and for some of them minor homology to characterised proteins was observed following BLAST analysis (Supplementary Table 1). Such hypothetical proteins may form the basis for future studies since no function (in platelets) could be assigned to them as yet.

4 Concluding remarks

In this study, we have used three versions of the COFRADIC technology to enrich three types of peptides during peptide-centric proteome analysis. By doing this, we first illustrate the versatility of COFRADIC, selecting *à la carte*, for either methionyl, cysteinyl or amino terminal peptides, while secondly, by combining the data generated by these approaches, we create the so far largest platelet proteome. It is interesting to notice that the set of proteins identified by each selective method is quite different with poor overlaps. As already mentioned before, this is most likely due to an undersampling phenomenon. Indeed, peptide-centric approaches are probably able to pick up most (even the minor) components of a mixture. However, because of the high flux of peptides on the one hand, and the large duty cycles of the mass spectrometer on the other hand, the latter only picks up a fraction of the passing peptide ions in a random manner with some preference for the most abundant peptides (proteins). Due to this random sampling process, it is necessary to carry out repetitive analyses in order to obtain a satisfactory level of protein coverage [26].

Comparison with the list of platelet proteins identified from previous 2-D gel studies [3, 4, 6, 24] reveals about 40% overlap with the proteins from our core list. Although this is without doubt much better than the low levels of overlap reported by similar comparative studies, it still reflects the undersampling effect mentioned above, creating a false image of complementarity between non-gel and 2-D gel approaches. Combining the information content of previously reported proteomes with those of the platelet core proteome generates a catalogue of over 1000 proteins which may form the basis from which different types of platelet research can initiate.

So far our core proteome contains 641 proteins. This is a high number, but clearly does not represent the full proteome of platelets. However, we have reasons to believe that a great part of the major platelet proteins has been covered. This is illustrated by the identification of all enzymes of the glycolytic and pentose phosphate pathways, and by allocating all eight subunits of the CCT chaperonin complex [35] and all seven components of the Arp2/3 complex [36].

Eighty-seven proteins in the core list contain at least one predicted transmembrane helix, 69 of these were previously characterised as membrane proteins. Only 12 of them were identified before in platelet proteomes. This illustrates the superiority of peptide-centric over protein-centric proteome approaches when it comes to identifying membrane proteins. This can be explained by the fact that hydrophobic proteins, which as a whole are hardly soluble, still may generate hydrophilic peptides which can be detected and serve as signatures for their parent proteins [37]. Even proteins with up to 7 and 12 predicted membrane-spanning helices and high GRAVY values were identified based upon a combination of different peptides.

High-throughput gel-free proteome analyses provide large sets of proteome data from different cell types. Although containing a wealth of information, these rather long lists of proteins at first appear to be but descriptive. However, comparing different proteomes using meta-data stored in for instance the Gene Ontology database indicates that COFRADIC identifies a core set of protein functions expressed by different human cell types (Fig. 3B). One may therefore expect that this core set will pop up in each proteome of cul-

Table 2. Identified platelet proteins that contain at least one helix spanning a biological membrane. Proteins are sorted according to the number of their known transmembrane helices (column 'KnownHel') (N.K. indicates that this number is not known). Such helices were predicted with the TMHMM Server v. 2.0 and these results are given in the column 'PredHel'. GRAVY values of the proteins were determined with the ProtParam tool at http://www.expasy.org/tools/protparam.html; positive GRAVY values indicate that the identified proteins are hydrophobic, negative values indicate hydrophilic proteins. Entries highlighted in a greyish background correspond to proteins that have previously been characterised in previous platelet proteome studies (see text)

	IPI ID	Description	KnownHel	PredHel	GRAVY value
1	IPI00337541.2	NAD(P) transhydrogenase, mitochondrial precursor	14	12	0.297
2	IPI00003909.1	Solute carrier family 2, facilitated glucose transporter, member 3	12	10	0.536
3	IPI00177817.4	Sarcoplasmic/endoplasmic reticulum calcium ATPase 2 (splice isoform SERCA2A)	10	7	0.087
4	IPI00004092.2	Sarcoplasmic/endoplasmic reticulum calcium ATPase 3 (splice isoform SERCA3B)	10	7	0.077
5	IPI00302840.1	Sodium/potassium-transporting ATPase alpha-3 chain	10	10	−0.016
6	IPI00151710.6	Similar to RIKEN cDNA F730003B03	8	8	−0.147
7	IPI00027180.1	CAAX prenyl protease 1 homolog	7	7	0.119
8	IPI00028332.1	Taste receptor type 2 member 13	7	7	0.819
9	IPI00032150.1	Phosphatidate cytidylyltransferase 2	6	8	0.275
10	IPI00152536.2	Transmembrane cochlear-expressed protein 2	6	9	−0.356
11	IPI00023639.1	CDP-diacylglycerol – inositol 3-phosphatidyltransferase	5	4	0.613
12	IPI00025292.1	Mannose-P-dolichol utilization defect 1 protein	5	5	0.679
13	IPI00215998.1	CD63 antigen	4	4	0.767
14	IPI00020446.1	CD82 antigen	4	4	0.378
15	IPI00375373.1	Chemokine-like factor super family member 5 (splice isoform 1)	4	4	0.448
16	IPI00218850.4	Secretory carrier-associated membrane protein 2	4	4	0.135
17	IPI00000612.1	Small membrane protein 1	4	4	0.580
18	IPI00218200.3	B-cell receptor-associated protein 31	3	3	−0.157
19	IPI00032038.3	Carnitine O-palmitoyltransferase I, mitochondrial liver isoform	2	2	−0.267
20	IPI00418495.1	CD36 antigen	2	2	0.015
21	IPI00017510.1	Cytochrome c oxidase polypeptide II	2	2	0.460
22	IPI00033075.1	Protein BAT5	2	2	−0.243
23	IPI00021766.3	Reticulon 4 (splice isoform 1)	2	2	−0.414
24	IPI00013897.1	ADAM 10 precursor	1	1	−0.599
25	IPI00006608.1	Amyloid beta A4 protein precursor (splice isoform APP770)	1	1	−0.584
26	IPI00019967.1	Apoptosis regulator BAX, membrane isoform alpha	1	1	−0.055
27	IPI00020984.1	Calnexin precursor	1	1	−0.874
28	IPI00008578.1	CD226 antigen precursor	1	1	−0.225
29	IPI00301271.3	Dolichyl-diphosphooligosaccharide – protein glycosyltransferase 63 kDa subunit precursor	1	4	0.081
30	IPI00025874.1	Dolichyl-diphosphooligosaccharide – protein glycosyltransferase 67 kDa subunit precursor	1	1	−0.208
31	IPI00032003.1	Emerin	1	1	−0.716
32	IPI00216758.1	Endothelin-converting enzyme 1 (splice isoform A)	1	1	−0.385
33	IPI00026530.2	ERGIC-53 protein precursor	1	1	−0.542
34	IPI00144014.1	HLA class I histocompatibility antigen, Cw-7 alpha chain precursor	1	1	−0.536
35	IPI00029046.1	Hypothetical protein KIAA0152	1	1	−0.156
36	IPI00216221.1	Integrin alpha-6 precursor (splice isoform Alpha-6X1A)	1	1	−0.381
37	IPI00295976.3	Integrin alpha-IIb precursor (splice isoform 1)	1	1	−0.106
38	IPI00218629.1	Integrin alpha-IIb precursor (splice isoform 3)	1	1	−0.150
39	IPI00009465.1	Integrin beta-1 precursor (splice isoform beta-1A)	1	1	−0.407
40	IPI00303283.1	Integrin beta-3 precursor (splice isoform beta-3A)	1	1	−0.334
41	IPI00001754.1	Junctional adhesion molecule 1 precursor	1	2	−0.092
42	IPI00002334.1	Neuron specific protein family member 1	1	1	−0.295
43	IPI00011255.1	Platelet glycoprotein Ib alpha chain precursor	1	1	−0.152
44	IPI00007723.1	Platelet glycoprotein Ib beta chain precursor	1	1	0.311
45	IPI00027502.1	Platelet glycoprotein IX precursor	1	1	0.284
46	IPI00027410.1	Platelet glycoprotein V precursor	1	1	0.131
47	IPI00157687.1	Platelet/endothelial cell adhesion molecule	1	1	−0.325
48	IPI00294472.3	Protein CGI-100 precursor	1	2	−0.066
49	IPI00006072.1	Protein transport protein SEC61 gamma subunit	1	1	0.321
50	IPI00031421.3	Protein-tyrosine sulfotransferase 2	1	1	−0.094
51	IPI00295339.2	P-selectin precursor	1	1	−0.291

Table 2. Continued

	IPI ID	Description	KnownHel	PredHel	GRAVY value
52	IPI00023807.1	Semaphorin 4D precursor	1	2	−0.274
53	IPI00219682.2	Stomatin isoform a	1	1	0.043
54	IPI00329332.1	Syntaxin 12	1	1	−0.580
55	IPI00289876.1	Syntaxin 7	1	1	−0.631
56	IPI00028055.1	Transmembrane protein Tmp21 precursor	1	2	−0.171
57	IPI00006865.1	Vesicle trafficking protein SEC22b	1	1	−0.182
58	IPI00019982.5	Vesicle-associated membrane protein 3	1	1	−0.079
59	IPI00006211.1	Vesicle-associated membrane protein-associated protein B/C (splice isoform 1)	1	1	−0.380
60	IPI00009950.1	Vesicular integral-membrane protein VIP36 precursor	1	1	−0.364
61	IPI00152377.1	Source of immunodominant MHC-associated peptides	N.K.	10	0.038
62	IPI00220300.1	ATP synthase H+ transporting mitochondrial F0 complex, sununit F, isoform 2	N.K.	1	−0.169
63	IPI00216583.1	G6B-A protein precursor	N.K.	1	−0.187
64	IPI00107540.1	Hypothetical protein (BLAST: similar to Integrin beta-1 precursor)	N.K.	1	−0.404
65	IPI00295313.1	JAW1-related protein MRVI1A long isoform	N.K.	1	−0.697
66	IPI00005202.1	Membrane associated progesterone receptor component 2	N.K.	1	−0.493
67	IPI00009976.1	Putative T1/ST2 receptor binding protein precursor	N.K.	1	−0.005
68	IPI00060144.1	Similar to RIKEN cDNA 9430029K10 gene	N.K.	4	0.282
69	IPI00386818.1	Vesicle-associated soluble NSF attachment protein receptor	N.K.	1	−0.678

Table 3. List of protein kinases and protein phosphatases identified in our non-gel human platelet proteome. IPI ID numbers of the proteins are indicated in the second column and a brief protein description is given in the last column. Proteins that have already been identified following large-scale platelet proteome studies are given in a grey background

	Protein kinases	
1	IPI00011891.2	5′-AMP-activated protein kinase, catalytic alpha-1 chain
2	IPI00016791.1	cGMP-dependent protein kinase 1, alpha isozyme
3	IPI00024403.1	Copine III
4	IPI00005689.2	Hypothetical protein KIAA1124 (BLAST: weakly similar to Rho-associated protein kinase 1)
5	IPI00002232.1	Hypothetical protein KIAA1361 (BLAST: protein is weakly similar to TRAF2 and NCK interacting kinase)
6	IPI00013219.1	Integrin-linked protein kinase 1
7	IPI00003479.1	Mitogen-activated protein kinase 1
8	IPI00010466.2	Protein kinase C, beta type (splice isoform beta-I)
9	IPI00329236.2	Protein kinase C, delta type
10	IPI00307155.4	Rho-associated protein kinase 2
11	IPI00099756.1	Serine/threonine protein kinase 13 (splice isoform 1)
12	IPI00329211.2	Serine/threonine protein kinase 3
13	IPI00021156.4	Serine/threonine-protein kinase Duet
14	IPI00219447.6	Serine/threonine-protein kinase PAK 2
15	IPI00328867.1	SRC protein
16	IPI00179357.1	Titin
17	IPI00145805.1	TRAF2 and NCK interacting kinase (splice isoform 1)
18	IPI00215778.1	Tyrosine-protein kinase LYN (splice isoform LYN B)
19	IPI00218278.1	Tyrosine-protein kinase SYK (splice isoform short)
	Protein phosphatases	
1	IPI00233255.3	Protein-tyrosine phosphatase, non-receptor type 11
2	IPI00289082.1	Protein-tyrosine phosphatase, non-receptor type 12
3	IPI00395552.1	Protein-tyrosine phosphatase, non-receptor type 6 (splice isoform 1)
4	IPI00179415.3	Serine/threonine protein phosphatase 2B catalytic subunit, alpha isoform
5	IPI00218236.3	Serine/threonine protein phosphatase PP1-beta catalytic subunit

tivated human/mammalian cell lines and that the underlying proteins may differ between different cell types or tissues and may thus be considered as future biomarkers. Again, this highlights the value of peptide-centric proteomics in unravelling systems biology.

K.G. is a Postdoctoral Fellow and L.M a Research Assistant of the Fund for Scientific Research – Flanders (Belgium) (F.W.O. – Vlaanderen). The project was supported by research grants from the Fund for Scientific Research – Flanders (Belgium) (Project number G.0008.03), the Inter University Attraction Poles (IUAP, Project number P5/05), the GBOU-research initiative (Project number 20204) of the Flanders Institute of Science and Technology (IWT) and the European Union Interaction Proteome (sixth Framework Program).

5 References

[1] Gawaz, M., *Blood Platelets*, Georg Thieme Verlag, Stuttgart 2001, pp. 1–3.

[2] Nurden, A. T., Nurden, P., in: Michelson, A. D. (Ed.), *Platelets*, Academic Press, New York 2002, pp. 681–700.

[3] Marcus, K., Immler, D., Sternberger, J., Meyer, H. E., *Electrophoresis* 2000, *21*, 2622–2236.

[4] O'Neill, E. E., Brock, C. J., von Kriegsheim, A. F., Pearce, A. C *et al. Proteomics* 2002, *2*, 288–305.

[5] Coppinger, J. A., Cagney, G., Toomey, S., Kislinger, T. *et al.*, *Blood* 2004, *103*, 2096–2104.

[6] Garcia, A., Prabhakar, S., Brock, C. J., Pearce, A. C. *et al.*, *Proteomics* 2004, *4*, 656–668.

[7] McRedmond, J. P., Park, S. D., Reilly, D. F., Coppinger, J. A. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 133–144.

[8] Zhang, H., Yan, W., Aebersold, R., *Curr. Opin. Chem. Biol.* 2004, *8*, 66–75.

[9] Brown, J. R., Hartley, B. S., *Biochem. J.* 1966, *101*, 214–228.

[10] Cruickshank, W. H., Malchy, B. L., Kaplan, H., *Can. J. Biochem.* 1974, *52*, 1013–1017.

[11] Gevaert, K., Van Damme, J., Goethals, M., Thomas, G. R. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 896–903.

[12] Gevaert, K., Ghesquière, B., Staes, A., Martens, L. *et al.*, *Proteomics* 2004, *4*, 897–908.

[13] Gevaert, K., Goethals, M., Martens, L., Van Damme, J. *et al.*, *Nat. Biotechnol.* 2003, *21*, 566–569.

[14] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[15] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y. *et al.*, *Proteomics* 2004, *4*, 1985–1988.

[16] Lagerwerf, F. M., van de Weert, M., Heerma, W., Haverkamp, J., *Rapid Commun. Mass Spectrom.* 1996, *10*, 1905–1910.

[17] Spengler, B., Kirsch, D., Kaufmann, R., *Rapid Commun. Mass Spectrom.* 1991, *5*, 198–202.

[18] Hermanson, G. T., *Bioconjugate Techniques*, Academic Press, London 1996, pp. 56–57.

[19] Roepstorff, P., Fohlman, J., *Biomed. Mass Spectrom.* 1984, *11*, 601.

[20] Washburn, M. P., Wolters, D., Yates, J. R. 3rd, *Nat. Biotechnol.* 2001, *19*, 242–247.

[21] Veenstra, T. D., Conrads, T. P., Issaq, H. J., *Electrophoresis* 2004, *25*, 1278–1279.

[22] Hardwidge, P. R., Rodriguez-Escudero, I., Goode, D., Donohoe, S. *et al.*, *J. Biol. Chem.* 2004, *279*, 20127–20136.

[23] Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., Gygi, S. P., *Proc. Natl. Acad. Sci. USA* 2003, *100*, 6940–6945.

[24] SWISS-2DPAGE release 17.1 at http://www.expasy.org/ch2d/.

[25] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D. *et al.*, *Nat. Genet.* 2000, *25*, 25–29.

[26] Ducret, A., Van Oostveen, I., Eng, J. K., Yates, J. R. 3rd, Aebersold, R., *Protein Sci.* 1998, *7*, 706–719.

[27] Liu, H., Sadygov, R. G., Yates, J. R. 3rd, *Anal. Chem.* 2004, *76*, 4193–4201.

[28] Kakhniashvili, D. G., Bulla, L. A. Jr., Goodman, S. R., *Mol. Cell. Proteomics* 2004, *3*, 501–509.

[29] Rabilloud, T., Adessi, C., Giraudel, A., Lunardi, J., *Electrophoresis* 1997, *18*, 307–316.

[30] Vuillard, L., Braun-Breton, C., Rabilloud, T., *Biochem. J.* 1995, *305*, 337–343.

[31] Moller, S., Croning, M. D., Apweiler, R., *Bioinformatics* 2001, *17*, 646–653.

[32] Kyte, J., Doolittle, R. F., *J. Mol. Biol.* 1982, *157*, 105–132.

[33] Jackson, S. P., Schoenwaelder, S. M., *Nat. Rev. Drug Discov.* 2003, *2*, 775–789.

[34] Marcus, K., Moebius, J., Meyer, H. E., *Anal. Bioanal. Chem.* 2003, *376*, 973–993.

[35] Rommelaere, H., Van Troys, M., Gao, Y., Melki, R. *et al.*, *Proc. Natl. Acad. Sci. USA* 1993, *90*, 11975–11979.

[36] Welch, M. D., Iwamatsu, A., Mitchison, T. J., *Nature* 1997, *385*, 265–269.

[37] Wu, C. C., MacCoss, M. J., Howell, K. E., Yates, J. R. 3rd, *Nat. Biotechnol.* 2003, *21*, 532–538.

3. Conclusions

3.1. Three goals for high-throughput proteomics

The development of dramatically improved instrumentation and novel techniques that maximize the capabilities of these instruments has delivered the promise of high-throughput proteomics in only a few years. The amounts of data produced rose in lockstep, with data yields rising several orders of magnitude over the same period. Interestingly, the most powerful new proteomics techniques are aimed primarily at maximizing the information content of the recorded data. This has the paradoxical effect of reducing the amount of redundancy in the generated data, thus prompting the need to make efficient use of the source information obtained. As the data finally obtained is often of high value to the scientific community it is often published as supplementary information but is renderred almost completely inaccessible through the use of PDF tables.

There are therefore three goals to be defined for any such high-throughput proteomics laboratory: managing the information flood, maximizing the value of the generated information and allowing researchers to share their information with the community at large. Finally, it is important to note these three goals are interdependent as each builds upon the achievement of the previous one.

There are also some issues to resolve for the field of proteomics as a whole. The first of these centers around its reluctance to take up an active role as a data producer with regards to sequence database providers. The second has to do with the software that is written for application in proteomics research laboratories and how it is made open source.

3.2. Managing information

In order to manage the thousands of fragmentation spectra produced daily by the different mass spectrometers in the proteomics laboratory at Ghent University, the ms_lims system was built. This system was designed to present a flexible, stable and user-friendly environment for the experimentalists and data analysts alike while building on low-cost, of-the-shelf components running open, freely available software components. The software has already gone through more than eighty released versions and continues to be updated today. The result is a highly mature, maintained software suite with a robust set of tools that support easy upgrading of both data stores and code base. Much of the code also operates in a frameworked design that allows third-party developers to add their own functionalities as required without having to rebuild the underlying software infrastructure from scratch. ms_lims has been freely available under the GNU GPL license since its inception and has been installed and tested in several locations. It has automated the complete identification pipeline, reducing processing time for a single proteome from weeks to one or two days. The time gained through the use of ms_lims can then be used to further analyse the final results, producing additional value from the data. As the stored data is maintained effortlessly over time, reanalysis of a given dataset is immediately possible at any point and cross-comparisons among disparate datasets are

suddenly possible. The ms_lims suite can thus be safely called the information backbone of the proteomics laboratory, with an analytical nervous system attached.
In summary, the goal of the ms_lims system was to automate routine tasks in order to speed up data processing time and reduce the possibility of human error while aiding the experimentalists or biologists in obtaining structured overviews of the relevant data for their research, thus maximizing their analytic capabilities.

3.3. Maximizing information

The COFRADIC peptide-centric proteomics approach allows researchers to maximize the analytical capabilities of modern mass spectrometers by greatly limiting redundancy in the generated information. This enables the instruments to look deeper into the proteome while simultaneously broadening the analysed fraction of the proteome.
In order to extract the most value from the recorded fragmentation spectra, it is necessary to distinguish between those spectra that are of high quality and those that are not. High quality spectra can then be subjected to exhaustive analysis in order to maximize the amount of identifications, while low quality spectra can be discarded. The spectrum quality assignment can also aid in identifying suspect identifications, thus increasing the reliability of the remaining identifications. In order to perform this task, a spectrum classification algorithm was developed through collaboration between the Proteomics Unit of the University of Bergen (PROBE) in Norway and the proteomics laboratory in Ghent. This software allows the separation of spectra to high quality and low quality bins according to a preset level of stringency. It is therefore possible to use it as a pre-filtering algorithm at low stringency (allowing some low quality spectra to pass through the filter in order to maximize the correct classification of high quality spectra) and as an erroneous identification detection tool at high stringency after the identification step. The algorithm builds on a learning algorithm called a Bayesian classifier that can be trained to respond correctly to different types of instruments and different types of analyzed spectra. The classifier software and code have been made publicly available under the GNU GPL license and have also been published.
For exhaustive identification of good quality spectra, it is important to adapt existing identification software with proven reliability to peptide-centric principles. Additionally, popular protein sequence databases must simultaneously be transformed to a far richer and more efficient source of peptide information. Both effects are achieved by the processing of protein centric sequence databases into peptide sequence databases. The DBToolkit software was developed to fulfill this task in a user-friendly yet highly automatable way. DBToolkit was also built around several interconnected core frameworks to allow third-party developers to quickly expand the existing functionality. The application is open source under the GNU GPL license and has been published.

3.4. Sharing information with the community

Scientific research is never a solitary task. Although the path from the initial sample to a final list of annotated protein identifications may be a long and lonely one, all this hard work will typically culminate in a publication in a peer-reviewed journal. Since science is essentially a collaborative effort, most journals demand publication of the presented

results. In order to allow all interested parties to maximize the value of these datasets, the PRIDE system was developed in collaboration with the European Bioinformatics Institute in Hinxton, UK. PRIDE started out as a one-man project during a stay as a Marie-Curie fellow at the EBI and has grown in two years to a mature and stable project supported by many local installations and containing almost 200.000 protein identifications, supported by well over half a million identified peptides. PRIDE continues to grow every day and has attracted the interest of journals and funding agencies. Two full-time equivalents constitute the core of the PRIDE development team, with collaborators across the globe contributing internationalization, data submission pipelines and code fragments. PRIDE is a HUPO PSI standards compliant repository and has always been at the forefront to implement and support these standards as they emerge. Consisting of both structured and semi-structured data elements using ontologies, PRIDE presents a common core of required data while allowing virtually limitless annotation of a specific dataset.
Being completely built out of open source components and in-house developed code released under the Apache license, there is no limitation to the use and incorporation of PRIDE components in new or existing software, regardless of any commercial interests. Preconfigured to work with both high-end professional database software as well as publicly available solutions, PRIDE can be downloaded and locally configured as an in-house or institutional data repository. It can also be used on a larger scale for international collaborations, obviating the need for them to develop their own proprietary infrastructure.
PRIDE has received attention from editorials in journals such as Nature and Expert Reviews in Proteomics and promises to grow as more and more authors and journals adopt the basic premise of scientific research: to share achievements with the scientific community at large.

3.5. The need for expanding the unidirectional relation between proteomics labs and sequence database providers

The relation between proteomics labs and database providers has so far been unidirectional in its strictest sense. The database providers make the fruits of their labour available over the internet, the proteomics lab downloads these databases and performs identifications. In the best case, the database provider is ultimately credited with a URL and/or reference in any publications resulting from the proteomics work.
Of course, being primarily a sequence database consumer is not a crime. It is interesting however, that the proteomics community to date has not grasped the fact that they have become an important group of data producers. The community has also failed to appreciate that the data produced is not only useful from a biological point of view but also from the perspective of sequence database providers.
It has been discussed in this dissertation that scientists working in proteomics have often commented on the disappearance of identifications when reconsidering certain identifications after sufficient time has passed (see section 1.2.2.4). Interestingly, these lamentations are only partially justified. Sequence database providers typically try to achieve two goals: provide sequence information (optionally adorned with additional annotations) of high quality and provide as much sequence information as possible. These goals are obviously somewhat incompatible. UniProt actually expressly hints at the

inherent conceptual differences by providing two subsets: UniProt/SWISS-PROT (biased towards quality of data) and UniProt/TrEMBL (biased towards quantity of data). It has been explained in section 1.2.2.2 that, in order to make the distinction between pure predictions and those that carry some experimental evidence, the NCBI - on a different level - similarly splits RefSeq into NP and XP entries. It should be noted that the evidence to move an entry from an XP accession to an NP accession is currently overwhelmingly derived from transcriptional studies. There are clear downsides to this approach in the generation of protein sequence entries, however (see [Claverie 2005] for an interesting discussion). Based on its prominent position as a data producer[60], proteomics should by now have become an important source of evidence for the validation of protein sequence predictions, replacing the apparently inferior evidence line of transcriptional data[61]. The question then is *why* proteomics has not taken its rightful position as an important evidence provider. The answer is deceivingly simply: because the proteomics community today is overwhelmingly confined to being a data consumer. Since little or no data at all flows back to the sequence database providers, the information obtained by proteomics techniques can not be used as evidence here.
The follow-up question then becomes: why is it that proteomics data does not flow back to the database providers? Most of the answers have been outlined in section 3.4, but apart from the practical considerations outlined there, a more subjective and controversial statement will be added here. This statement can be summarized by stating that the field of proteomics needs to mature. Achieving maturity as a scientific field requires more than good techniques, the acquisition of expensive instruments and long lists of protein identifications; it requires the practitioners to take responsibility for their produced data by expending the extra effort after the completion of a project to close the information sharing cycle all the way back to the initial database provider.

3.6. The case for open source software in the life sciences

All the software discussed in this dissertation has been made freely available as open source software. Licenses are either the non-restrictive Apache license or the 'infective' GNU GPL license. The software discussed has also been developed to be readily extensible and contains ample documentation for the aspiring contributor.
These commitments to provide the source code are not without their price. Code opened to the community should be well documented, which is a common flaw of computer programmers. It should also be made available through an anonymous, open infrastructure such as the Concurrent Versioning System (CVS) or SubVersion and it should be able to accept code input from (trusted) third-party developers. Additionally, it should also be possible to use the software as a module in a third-party application, which in turn requires a certain amount of stability and a strict and traceable versioning of the individual source files and the whole project alike. Careful logging of each modification

[60] Also consider the wealth of additional information available through positional proteomics techniques such as N-terminal COFRADIC!
[61] It should be stressed that the adjective 'inferior' is here applied for *protein* sequence predictions only. Transcriptional data is obviously pretty good evidence that certain genomic sequences are converted into RNA.

– especially where it concerns key algorithmic or API changes – is another major requirement for software upon which other people depend.
All of the above are considered normal and are therefore very much standard practice in the software development industry. The lone hacker working in a mid-size proteomics lab or the mathematician providing a ready application to test her new algorithm might not feel up to the challenges involved in setting up all this infrastructure and code and project lifecycle management. However, when one looks at the above checklist from a purely objective point of view, it becomes clear that these (ultimately) simple and straightforward rules are essential to allowing others to productively benefit from one's own development efforts. Open source requirements in proteomics should therefore be more than a request by journals to provide a zip file with some text documents (the code) on a website for a few months. Open source should instead become a way of working that is steeped in those practices that allow true collaborative efforts to emerge.

4. Nederlandstalige samenvatting

Het veld van de proteoomanalyse is een snel groeiende discipline die toepassingen kent in fundamenteel en toegepast wetenschappelijk onderzoek. Voortbouwend op de indrukwekkende verwezenlijkingen van de genoomsequeneringsprojecten en een sterke technologische vooruitgang van de instrumenten, heeft de proteoomanalyse de laatste jaren inderdaad een enorme vooruitgang gekend. De traditionele methode van de twee-dimensionele poly-acrylamide gelelectroforese voor eiwitscheiding gevolgd door eiwitidentificatie door middel van massaspectrometrische data, sequentiedatabanken en zoekalgoritmen is dan ook naar een zeer courant gebruikte techniek geëvolueerd. Deze techniek, hoewel erg succesvol, is echter onderhevig aan enkele belangrijke limitaties. De belangrijkste hiervan betreffen hydrofobe eiwitten en laag abundante eiwitten. De hydrofobe eiwitten zijn erg moeilijk in oplossing te houden en ontsnappen dan ook erg gemakkelijk aan identificatie omwille van precipitatie. De laag abundante eiwitten ontsnappen op hun beurt aan verdere analyse omdat zij niet gevisualiseerd kunnen worden na scheiding.

Deze problemen worden echter aangepakt door verschillende nieuwe methoden voor proteoomanalyse. Deze nieuwe technieken onderscheiden zich van de klassieke gel-gebaseerde methode door een verschuiving in de eenheid van analyse. Deze verschoof van het eiwit naar de peptiden die verkregen worden door middel van proteolytische digestie. Aangezien de peptiden een veel minder breed bereik van fysico-chemische parameters vertonen, zijn ze gemakkelijker te scheiden en in oplossing te houden. Bovendien levert een enkel eiwit gemiddeld een dertigtal verschillende peptiden wanneer het met het enzyme trypsine wordt behandeld. Deze redundantie zorgt ervoor dat voor sommige van deze peptide-gecentreerde technieken het missen van een bepaalde peptide gemakkelijk gecompenseerd kan worden door de vele andere peptiden die van hetzelfde eiwit afkomstig zijn.

De toegenomen redundantie brengt echter ook een overeenkomstige toename in complexiteit met zich mee. In combinatie met het brede concentratiebereik van eiwitten in cellen zorgt deze complexiteit dikwijls voor problemen bij die peptide-centrische technieken die het volledige peptidenmengsel over één of meerdere dimensies (bv. ionenuitwisselingschromatografie en omgekeerde-fase hogedrukschromatografie) scheiden. De scheiding in de tijd van de peptiden is immers nooit perfect, waardoor er steeds verschillende peptiden in overlappende tijdsintervallen zullen elueren. Dit stelt een probleem voor de massaspectrometer aangezien deze een keuze moet maken in verband met het peptide dat geselecteerd zal worden voor gedetailleerde analyse door fragmentatie. Gedurende deze fragmentatiestap is de massaspectrometer blind voor andere peptiden die elueren. Wanneer nu verschillende peptiden tegelijkertijd aangeboden worden aan de massaspectrometer, zal er dus slechts een beperkt aantal in detail geanalyseerd kunnen worden. In de praktijk komt het er vaak op neer dat peptiden afkomstig van abundante eiwitten meer kans maken om geanalyseerd te worden.

Om dit probleem te vermijden werden er andere peptide-gecentreerde technieken ontwikkeld die eerst de complexiteit van het mengsel na digestie trachten te reduceren. Het is belangrijk in te zien dat deze stap tegelijkertijd de representativiteit van het mengsel zo goed mogelijk dient te behouden. Een van deze selectiemethoden werd ontwikkeld in de proteoomanalytische groep van Prof. Dr. Joël Vandekerckhove en is

gebaseerd op de reeds lang gekende techniek van de diagonaalchromatografie. Deze techniek kreeg het acronym COFRADIC (voor *COmbined FRActional DIagonal Chromatography*) toebedeeld. Hierbij wordt een mengsel tweemaal onderworpen aan een identieke chromatografische stap, maar introduceert men tussen de eerste en de tweede stap een specifieke wijziging aan bepaalde peptiden waardoor hun elutietijd wijzigt. In de tweede chromatografische stap zullen deze peptiden dan herkenbaar zijn aan hun veranderde elutietijd. De wijzing aan de te selecteren peptiden kan bereikt worden door eender welke chemische of enzymatische reactie (met het voorbehoud dat deze reactie *specifiek* moet zijn), hetgeen een zeer brede waaier aan mogelijke selectiecriteria toelaat. Zo kan men selecteren op basis van peptidensequentie (bv. alle methionyl of cysteinyl peptiden), het dragen van post-translationele modificaties (bv. fosforylatie) of de lokalisatie van een peptide binnen een eiwit (bv. het aminoterminale peptide).
Gezien de verschuiving van de analytische eenheid naar het peptide bij deze nieuwe proteoomanalytische methoden, is het nodig ook de gegevensanalyse aan te passen. Zeker in het geval van de selectiemethoden zoals COFRADIC is het belangrijk om rekening te houden met deze evolutie. Inderdaad, wanneer enkel bepaalde peptiden geselecteerd worden uit een complex mengsel, verdwijnt het voordeel van de redundantie tesamen met de complexiteit. Het wordt dan ook essentieel om elk peptide op te pikken aangezien het missen van een peptide kan gelijk staan aan het missen van een eiwit (bv. wanneer voor aminoterminale peptiden geselecteerd wordt; elk eiwit heeft immers slechts één aminoterminus).
Een eerste stap die nodig is om de beschikbare zoekalgoritmen en sequentiedatabanken meer efficiënt aan te kunnen wenden, betreft het omzetten van de eiwitgebaseerde databanken naar peptide-gebaseerde versies. Hiertoe werd een software applicatie ontwikkeld (*DBToolkit* genaamd) die de meest courante databankformaten kan inlezen en op verschillende manieren kan omzetten naar peptide-databanken. Hierbij zijn twee processen van belang: het extraheren van additionele informatie uit de databanken en het elimineren van de redundantie op peptidenniveau in de databanken. Dit eerste process is opnieuw voornamelijk belangrijk in het geval van aminoterminale COFRADIC. Vele eiwitten worden immers verknipt aan hun aminoterminus tijdens hun maturatieproces en deze knipplaatsen komen doorgaans niet overeen met de knipplaatsen van standaard proteolytische enzymes, gebruikt in de proteoomanalyse. Het gevolg is dat de zoekalgoritmen de uiteindelijke aminoterminale peptiden niet in hun lijst van potentiële peptiden opnemen wanneer zij een *in silico* proteolytisch digest van een eiwitsequentiedatabank maken. DBToolkit lost dit probleem op door van elk correct proteolytisch peptide een collectie aminoterminaal ‘gerafelde’ peptiden te maken. Deze peptiden vormen samen een lijst van alle mogelijke aminoterminale peptiden van de eiwitten in de sequentiedatabank, ongeacht de locatie van de knipplaats. De tweede bewerking, het verwijderen van de redundantie op peptidenniveau, is een meer algemeen geldende maatregel. De meeste sequentiedatabanken zijn niet-redundant op eiwitsequentie niveau. Dit houdt in dat elke volledige sequentie slechts éénmaal voorkomt in de databank. Dit is echter geen garantie voor het niet-redundant zijn van de peptidensequenties. Deze hangt in de praktijk sterk of van de gebruikte databank, maar loopt al snel op tot 50% van de peptidensequenties voor een tryptisch digest! DBToolkit pakt dit probleem aan door elke peptidensequenties slechts éénmaal te vermelden in de uiteindelijke databank maar dan wel met expliciete vermelding van alle eiwitten die

aanleiding kunnen geven tot dit peptide. Op die manier gaat er geen informatie verloren bij deze conversie en kan men spreken van een verliesloze compressietechniek. Het belang van deze operatie wordt duidelijk wanneer we bekijken hoe bepaalde zoekalgoritmen onderscheid maken tussen correct positieve identificaties en vals positieve identificaties. Deze algoritmen, die probabilistische algoritmen genoemd worden, berekenen een *a priori* drempelscore waarboven een identificatie moet scoren om – met een bepaalde betrouwbaarheid – als geïdentificeerd te worden geklasseerd. Deze drempelwaarde is afhankelijk van de grootte van de databank, zoals gemeten aan de hand van het aantal aminozuren in die databank. Het is duidelijk dat een databank met 50% redundantie dus dubbel zo groot wordt ingeschat als ze eigenlijk is. Dit resulteert in een artificieel hoge drempelwaarde en dit heeft een negatieve invloed op de identificatie-efficiëntie.

Een tweede probleem waarmee de nieuwe peptide-gecentreerde technieken mee geconfronteerd worden, betreft de hoeveelheid informatie die routinematig gegenereerd wordt. De datastromen zijn dan ook met enkele grootte-ordes toegenomen over de laatste jaren. Daar waar het enkele jaren geleden nog gebruikelijk was om met een rekenblad zoals Excel de identificaties voor een bepaald staal te beheren, is dit nu eenvoudigweg onmogelijk geworden. Omdat deze technieken jong en vernieuwend zijn, bestaat er geen degelijke commerciële software om dit gegevensbeheer te verzorgen. Het is daarom noodzakelijk om zelf een databank en bijhorende software te ontwikkelen. In dit werk wordt dan ook de ms_lims software besproken, die ontwikkeld werd om de datastromen bij de peptide-gecentreerde proteoomanalyse te automatiseren. Centraal in het systeem staat een relationele databank waarin de gegevens opgeslaan worden, met daarrond een set van applicaties die een eenvoudige interactie met de data mogelijk maken. Deze modulair opgebouwde software voldoet aan enkele belangrijke eisen. Zo integreert ze op een volledig transparante manier de informatie afkomstig van diverse instrumenten, hierbij zorgvuldig tracerend welke data van welk instrument komt. Er is ook een koppeling met een van de meest gebruikte zoekalgoritmen (Mascot) waardoor de resultaten van de zoekopdrachten volautomatisch in de relationele databank worden opgeslaan. Verder zijn er verschillende programma's voorhanden die een specifieke analyse van de gegevens toelaten. Een mooi voorbeeld hiervan is de applicatie die een differentiële analyse van één of meer projecten toelaat. Hierbij wordt er een, op robuuste statistiek gebaseerde, statistische analyse uitgevoerd op de verzamelde identificaties en kan de gebruiker eenvoudig de peptiden of eiwitten bestuderen die significant gereguleerd worden tussen verschillende stalen.

De data die door de massaspectrometers aangeleverd wordt, kan echter ook nog op een andere manier geanalyseerd worden. Deze massaspectra zijn immers niet steeds afkomstig van peptiden. Dikwijls heeft de massaspectrometer een polymeer of contaminant gefragmenteerd en hiervan een spectrum opgesteld. Het is uiteraard zinloos om zulke spectra in de volledige identificatie-cyclus mee te nemen daar er nooit een peptidesequentie aan gekoppeld kan worden. Naast deze *a priori* filtering kan men echter ook een *a posteriori* filtering doorvoeren waarbij na identificatie een onafhankelijke kwaliteitsscore wordt toegekend aan een spectrum en deze tweede score in ogenschouw wordt genomen bij het aanvaarden of verwerpen van een identificatie. Op deze manier kan dan gezocht worden naar de vals positieve identificaties die resteren binnen het betrouwbaarheidsinterval dat in het zoekalgoritme ingesteld werd. Het is duidelijk dat een

andere stringentie gewenst is bij het *a priori* filteren dan bij het *a posteriori* filteren. Wanner er immers een te strenge selectie doorgevoerd wordt vooraleer tot identificatie over te gaan, is er een grote kans dat de filter, naast een heleboel slechte spectra, ook een aanzienlijk aantal goede spectra zal wegfilteren. Aangezien het nu zaak is zo efficiënt mogelijk met de opgenomen spectra om te gaan, is het belangrijk om tijdens deze stap te stringente selectie te vermijden. Omgekeerd dient de selectie erg streng te gebeuren tijdens het valideren van de bekomen identificaties. Hier is het immers wenselijk elk verdacht spectrum te signaleren. Om deze filtering te kunnen uitvoeren, werd in samenwerking met Kristian Flikka en Professor Ingvar Eidhammer van de Universiteit Bergen in Noorwegen een applicatie ontwikkeld om spectra te filteren met een vrij te kiezen stringentie. Deze aanpak laat ook toe om die spectra op te pikken die een erg hoge kwaliteitsscore krijgen, maar toch niet geïdentificeerd worden. Deze spectra kunnen onderworpen worden aan meer arbeidsintensieve analyses om ze vooralsnog te kunnen identificeren.

Op een ander niveau worstelt de wereld van de proteoomanalyse met het probleem van het verspreiden en ter beschikking stellen van de bekomen identificaties. Tot voor kort werd dit soort data voornamelijk via onhandige en vaak incomplete PDF tabellen als supplementaire informatie verspreid door de wetenschappelijke vakbladen. Dit met enkele kwalijke gevolgen: spectra zijn niet inbegrepen in deze informatie zodat een herevaluatie en eventuele validatie van de bekomen identificaties onmogelijk wordt. Bovendien is de data in PDF formaat wel leesbaar voor mensen, maar niet voor computerprogramma's. Gezien de enorme hoeveelheid identificaties die typisch gepubliceerd worden, is automatische verwerking in de praktijk echter noodzakelijk. Het uiteindelijke resultaat van deze omstandigheden is dat de informatie, eenmaal gepubliceerd, vergeten wordt en niet kan bijdragen aan verder wetenschappelijk werk. Om al deze problemen aan te pakken, werd in samenwerking met het European Bioinformatics Institute (EBI) het PRIDE (PRoteomics IDEntifications) project ontwikkeld. Dit systeem bestaat uit een open bestandsformaat in XML voor de verspreiding van data, een centrale database waarin deze data kan worden opgeslaan en een volledige set van software bibliotheken om met databank en XML formaat te interageren. De centrale databank is via het web te raadplegen en kan eveneens gedownload worden via FTP. Bovendien is het volledige pakket inclusief de broncode vrij beschikbaar. Dit laat geïnteresseerden toe om het systeem lokaal te installeren en zelfs om eigen veranderingen aan te brengen. PRIDE laat ook toe om online data in te dienen. Deze gegevens kunnen voor (on-)bepaalde tijd privaat gehouden worden en bijvoorbeeld alleen gedeeld worden binnen samenwerkingsverbanden of met een editor en peer-reviewers.

Ten slotte wordt er een synthese-project voorgesteld: de toepassing van alle bovenvermelde applicaties en verschillende COFRADIC technieken om de meest complete dataset van het proteoom van humane bloedplaatjes samen te stellen.

In conclusie kan gesteld worden dat het veld van de proteoomanalyse stilaan volwassen wordt en dat specifieke software-applicaties hiertoe sterk hebben bijgedragen. De doorgedreven automatisering van de gegevensstroom laat toe om aan een zeer hoog tempo analyses uit te voeren en de grote hoeveelheden informatie die hieruit voortkomen, kunnen op hun beurt op een efficiënte manier geraadpleegd worden aan de hand van analyse-software. De proteoomanalyse zal in de volgende jaren wel nog inspanningen

moeten doen om de rol van data producent op een georganiseerde en nuttige manier te dragen. Momenteel wordt nog teveel data achtergehouden om verschillende, weinig wetenschappelijke redenen. De infrastructuurproblemen die hiervoor vroeger als excuus werden aangehaald, zijn inmiddels grotendeels opgelost en met de komst van centrale identificatie-databanken zoals PRIDE ligt de weg open naar een meer constructieve en coöperatieve manier van onderzoek doen.
Een laatste pleidooi betreft het overnemen van de methodes van de *open source* software-ontwikkeling. Hierbij wordt niet alleen de broncode op de een of andere manier vrijgegeven, maar wordt er ontwikkeld volgens welbepaalde methoden die het delen van de code, het gebruik van de software alsook het wereldwijd samenwerken van ontwikkelaars ondersteund en stimuleert.

5. References

Amft M, Moritz F, Weickhardt C, Grotemeyer J, 'Instrumental measures to enhance the mass resolution in matrix assisted laser desorption/ionization (MALDI) time-of-flight experiments: computational simulations and experimental observations', **1997**, *International Journal of Mass Spectrometry and Ion Processes* **167/168**:661-674.

Anderson DC, Li WQ, Payan DG, Noble WS, 'A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores', **2003**, *Journal of Proteome Research* **2**:137-146.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A *et al.,* 'Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes', **2002**, *Science* **297**:1301-1310.

Baczek T, Bucinski A, Ivanov AR, Kaliszan R, 'Artificial neural network analysis for evaluation of peptide MS/MS spectra in proteomics', **2004**, *Analytical Chemistry* **76**:1726-1732.

Beavis RC, Chait BT, 'Cinnamic acid derivatives as matrices for ultraviolet laser desorption mass spectrometry of proteins', **1989**, *Rapid Communications in Mass Spectrometry* **3**:432-435.

Beavis RC, Chait BT, 'High-accuracy molecular mass determination of proteins using matrix-assisted laser desorption mass spectrometry', **1990**, *Analytical Chemistry* **62**:1836-1840.

Biemann K, 'Contributions of mass-spectrometry to peptide and protein-structure', **1988**, *Biomedical & environmental mass spectrometry* **16**:99-111.

Bradshaw RA, 'Revised draft guidelines for proteomic data publication', **2005**, *Molecular and Cellular Proteomics* **4**:1223-1225.

Bronstrup M, 'Absolute quantification strategies in proteomics based on mass spectrometry', **2004**, *Expert Reviews in Proteomics* **1**:503-512.

Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A, 'The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data', **2004**, *Molecular and Cellular Proteomics* **3**:531-533.

Clauser KR, Baker PR, Burlingame AL, 'Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching', **1999**, *Analytical Chemistry* **71**:2871-2882.

Claverie JM, 'What If There Are Only 30,000 Human Genes?', **2001**, *Science*, **291**:1255-1257.

Claverie JM, 'Fewer genes, more noncoding RNA', **2005**, *Science* **309**:1529-1530.

Colinge J, Masselot A, Giron M, Dessingy T, Magnin J, 'OLAV: towards high-throughput tandem mass spectrometry data identification', **2003**, *Proteomics*, **3**:1454-1463.

Colinge J, Masselot A, Giron M, Dessingy T, Magnin J, 'OLAV: towards high-throughput tandem mass spectrometry data identification', **2003**, *Proteomics* **3**:1454-1463.

Collins FS, 'Contemplating the end of the beginning', **2001**, *Genome Research* **11**:641-643.

Cotter RJ, 'Reflectrons and other energy-focusing devices', **1997** In: **Time-of-Flight Mass Spectrometry: Instrumentation and Applications in Biological Research**, *ACS professional reference books*, Washington DC, pg. 47-72.

Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA, 'De novo peptide sequencing via tandem mass spectrometry', **1999**, *Journal of Computational Biology* **6**:327-342.

Denison C, Rudner AD, Gerber SA, Bakalarski CE, Moazed D, Gygi SP, 'A proteomic strategy for gaining insights into protein sumoylation in yeast', **2005**, *Molecular and Cellular Proteomics* **4**:246–254.

Eddes JS, Kapp EA, Frecklington DF, Connolly LM, Layton MJ, Moritz RL, Simpson RJ, 'CHOMPER: a bioinformatic tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies', **2002**, *Proteomics* **2**:1097-1103.

Elias JE, Haas W, Faherty BK, Gygi SP, 'Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations', **2005**, *Nature Methods* **2**:667-675.

Eng J, McCormack AL, Yates JR 3rd, 'An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database', **1994**, *Journal of the American Society for Mass Spectrometry* **5**:976-989.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM, 'Electrospray ionization for mass spectrometry of large biomolecules', **1989**, *Science* **246**:64-71.

Fenyo D, Beavis RC, 'A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes', **2003**, *Analytical Chemistry* **75**:768-774.

Fernandez-de-Cossio J, Gonzalez J, Satomi Y, Shima T, Okumura N, Besada V, Betancourt L, Padron G, Shimonishi Y, Takao T, 'Automated interpretation of low-energy collision-induced dissociation spectra by SeqMS, a software aid for de novo sequencing by tandem mass spectrometry', **2000**, *Electrophoresis* **21**:1694-1699.

Field HI, Fenyo D, Beavis RC, 'RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database', **2002**, *Proteomics* **2**:36-47.

Frank A, Pevzner PA, 'PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling', **2005**, *Analytical Chemistry* **77**:964-973.

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH, 'Open mass spectrometry search algorithm', **2004**, *Journal of Proteome Research*, **3**:958-964.

Gevaert K, Van Damme J, Goethals M, Thomas GR, Hoorelbeke B, Demol H, Martens L, Puype M, Staes A, Vandekerckhove J, 'Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis: identification of more than 800 Escherichia coli proteins', **2002**, *Molecular and Cellular Proteomics* **1**:896-903.

Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR, Vandekerckhove J, 'Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides', **2003**, *Nature Biotechnology* **21**:566-569.

Gevaert K, Ghesquiere B, Staes A, Martens L, Van Damme J, Thomas GR, Vandekerckhove J, 'Reversible labeling of cysteine-containing peptides allows their specific chromatographic isolation for non-gel proteome studies', **2004**, *Proteomics* **4**:897-908.

Gevaert K, Van Damme P, Martens L, Vandekerckhove J, 'Diagonal reverse-phase chromatography applications in peptide-centric proteomics: ahead of catalogue-omics?', **2005**, *Analytical Biochemistry* **345**:18-29.

Gevaert K, Staes A, Van Damme J, De Groot S, Hugelier K, Demol H, Martens L, Goethals M, Vandekerckhove J, 'Global phosphoproteome analysis on human HepG2 hepatocytes using reversed-phase diagonal LC', **2005**, *Proteomics* **5**:3589-3599.

Gevaert K, Pinxteren J, Demol H, Hugelier K, Staes A, Van Damme J, Martens L, Vandekerckhove J, ‘A four stage liquid chromatographic selection of methionyl peptides for peptide-centric proteome analysis: the proteome of human multipotent adult progenitor cells’, *Journal of Proteome Research submitted*.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE *et al.*, 'Genome sequence of the Brown Norway rat yields insights into mammalian evolution', **2004**, *Nature* **428**:493-521.

Görg A, Postel W, Gunther S, Weser J, Strahler JR, Hanash SM, Somerlot L, Kuick R, 'Approach to stationary two-dimensional pattern: influence of focusing time and immobiline/carrier ampholytes concentrations', **1988**, *Electrophoresis* **9**:37-46.
Görg A, Obermaier C, Boguth G, Csordas A, Diaz JJ, Madjar JJ 'Very alkaline immobilized pH gradients for two-dimensional electrophoresis of ribosomal and nuclear proteins', **1997**, *Electrophoresis*, **18**:328-337.

Görg A, Obermaier C, Boguth G, Weiss W, 'Recent developments in two-dimensional gel electrophoresis with immobilized pH gradients: wide pH gradients up to pH 12, longer separation distances and simplified procedures', **1999**, *Electrophoresis* **20**:712-717.

Grossmann J, Roos FF, Cieliebak M, Lipták Z, Mathis LK, Müller M, Gruissem W, Baginsky S, 'AUDENS: A Tool for Automated Peptide de Novo Sequencing', **2005**, *Journal of Proteome Research* **4**:1768-1774.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R, 'Quantitative analysis of complex protein mixtures using isotope-coded affinity tags', **1999**, *Nature Biotechnology* **17**:994-999.

Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R, 'Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology', **2000**, *Proceedings of the National Academy of Sciences U. S. A*. **15**:9390-9395.

Hampel FR, Introduction to "Huber (1964), Robust estimation of a location parameter" **Breakthroughs in Statistics, Volume II**, **1993**, *Springer*, New York.

Hanash S, Celis JE, 'The Human Proteome Organization: a mission to advance proteome knowledge', **2002**, *Molecular and Cellular Proteomics* **1**:413-414.

Harrison PM, Kumar A, Lang N, Snyder M and Gerstein M, 'A question of size: the eukaryotic proteome and the problems in defining it', **2002**, *Nucleic Acids Research*, **30**:1083-1090.

Hernandez P, Gras R, Frey J, Appel RD, 'Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data', **2003**, *Proteomics* **3**:870-878.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME *et al.,* 'Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution', **2004**, *Nature* **432**:695-716.

Huber PJ, **Robust statistics**, **1981**, *John Wiley and Sons Inc.*, New York.

International Human Genome Sequencing Consortium, 'Finishing the euchromatic sequence of the human genome', **2004**, *Nature* **431**:931-45.

Jonscher KR, Yates I, John R, 'The quadrupole ion trap mass spectrometer: a small solution to a big challenge', **1997**, *Analytical Biochemistry* **244**:1-15.

Jørgensen T, Bojesen G, Rahbek-Nielsen H, 'The proton affinities of seven matrix-assisted laser desorption/ionization matrices correlated with the formation of multiply charged ions', **1998**, *European Journal of Mass Spectrometry* **4**:39-45.

Julka S, Regnier FE, 'Recent advancements in differential proteomics based on stable isotope coding', **2005**, *Briefings in Functional Genomics and Proteomics* **4**:158-177.

Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ, 'An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis', **2005**, *Proteomics* **5**:3475-3490.

Karas M, Hillenkamp F, 'Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons', **1988**, *Analytical Chemistry* **60**:2299-2301.

Karas M, Bahr U, Stahl-Zeng JR, 'Steps towards a more refined picture of the matrix function in UV MALDI', **1996**, In: Baer, T., Ng, C.Y. and Powis, I. (eds.) **Large Ions: Their Vaporization, Detection and Structural Analysis**, *Wiley-VCH*, Weinheim, pg. 27-48.

Keller A, Nesvizhskii AI, Kolker E, Aebersold R, 'Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search', **2002**, *Analytical Chemistry* **74**:5383-5392.

Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R, 'The International Protein Index: an integrated database for proteomics experiments', **2004**, *Proteomics* **4**:1985-1988.

Khidekel N, Ficarro SB, Peters EC, Hsieh-Wilson LC, 'Exploring the O-GlcNAc proteome: direct identification of O-GlcNAc modified proteins from the brain', **2004**, *Proceedings of the National Academy of Sciences U. S. A.* **101**:13132–13137.

Kho Y, Kim SC, Jiang C, Barma D, Kwon SW, Cheng J, Jaunbergs J, Weinbaum C, Tamanoi F, Falck J, Zhao Y, 'A tagging-via-substrate technology for detection and proteomics of farnesylated proteins', **2004**, *Proceedings of the National Academy of Sciences U. S. A.* **101**:12479–12484.

Kirkpatrick DS, Gerber SA, Gygi SP, 'The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications', **2005**, *Methods* **35**:265-273.

Klose J, 'Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals', **1975**, *Humangenetik* **26**:231-243

Kristensen DB, Brond JC, Nielsen PA, Andersen JR, Sorensen OT, Jorgensen V, Budin K, Matthiesen J, Veno P, Jespersen HM *et al.*, 'Experimental Peptide Identification Repository (EPIR): an integrated peptide-centric platform for validation and mining of tandem mass spectrometry data', **2004**, *Molecular and Cellular Proteomics* **3**:1023-1038.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.,* 'Initial sequencing and analysis of the human genome', **2001**, *Nature*, **409**:860-921.

Li F, Sun W, Gao YH, Wang J, 'RScore: a peptide randomicity score for evaluating tandem mass spectra', **2004**, *Rapid Communications in Mass Spectrometry* **18**:1655-1659.

Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR 3rd, 'Direct analysis of protein complexes using mass spectrometry', **1999**, *Nature Biotechnology* **17**:676–682.

Liu H, Sadygov RG, Yates JR 3rd, 'A model for random sampling and estimation of relative protein abundance in shotgun proteomics', **2004**, *Analytical Chemistry* **76**:4193-4201.

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G, 'PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry', *2003*, *Rapid Communications in Mass Spectrometry* **17**:2337-2342.

Mamyrin BA, Karataev VI, Shmikk DV, Zagulin VA, 'The Mass-Reflectron, a New Nonmagnetic Time-of-Flight Mass Spectrometer with High Resolution', **1973**, *Soviet Physics-JETP* **37**:45–48.

Mann M, Hojrup P, Roepstorff P, 'Use of mass spectrometric molecular weight information to identify proteins in sequence databases', **1993**, *Biological Mass Spectrometry* **22**:338-345.

Mann M, Wilm M, 'Error-tolerant identification of peptides in sequence databases by peptide sequence tags', **1994**, *Analytical Chemistry* **66**:4390-4399.

March RE, 'An introduction to quadrupole ion trap mass spectrometry', **1996**, *Journal of Mass Spectrometry* **32**:351-369.

Marko-Varga G, Nilsson J, Laurell T, 'New directions of miniaturization within the proteomics research area', **2003**, *Electrophoresis* **24**:3521-3532.

McDonald WH, Tabb DL, Sadygov RG, MacCoss MJ, Venable J, Graumann J, Johnson JR, Cociorva D, Yates JR 3rd, 'MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications', **2004**, *Rapid Communications in Mass Spectrometry* **18**:2162-2168.

Meyer TS, Lamberts BL, 'Use of coomassie brilliant blue R250 for the electrophoresis of microgram quantities of parotid saliva proteins on acrylamide-gel strips', **1965**, *Biochimica et Biophysica Acta* **24**:144-145.

Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang SP, Enard W, Hellmann I, Lindblad-Toh K, Altheide TK *et al.*, 'Initial sequence of the chimpanzee genome and comparison with the human genome', **2005**, *Nature* **437**:69-87.

Monteoliva L, Albar JP, 'Differential proteomics: an overview of gel and non-gel based approaches', **2004**, *Briefings in Functional Genomics and proteomics* **3**:220-239.

Nesvizhskii AI, Keller A, Kolker E, Aebersold R, 'A statistical model for identifying proteins by tandem mass spectrometry', **2003**, *Analytical Chemistry* **75**:4646-4658.

Nesvizhskii AI, Aebersold R, 'Interpretation of shotgun proteomic data: the protein inference problem', **2005**, *Molecular and Cellular Proteomics* **4**:1419-1440.

Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R, 'Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides', **2005**, *Molecular and Cellular Proteomics e-pub ahead of print*.

Oda Y, Nagasu T, Chait BT, 'Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome', **2001**, *Nature Biotechnology* **19**:379-382.

O'Farrell PH, 'High resolution two-dimensional electrophoresis of proteins', **1975**, *Journal of Biological Chemistry* **250**:4007-4021

O'Farrell PZ, Goodman HM, O'Farrell PH, 'High resolution two-dimensional electrophoresis of basic as well as acidic proteins', **1977**, *Cell* **12**:1133-1141.

Olsen JV, Ong SE, Mann M, 'Trypsin cleaves exclusively C-terminal to arginine and lysine residues', **2004**, *Molecular and Cellular Proteomics* **3**:608-614.

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M, 'Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate

approach to expression proteomics', **2002**, *Molecular and Cellular Proteomics* **1**:376–386.

Orchard S, Hermjakob H, Taylor CF, Potthast F, Jones P, Zhu W, Julian RK Jr, Apweiler R, 'Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17-20th April 2005)', **2005**, *Proteomics* **5**:3552-3555.

Pappin DJ, Hojrup P, Bleasby AJ, 'Rapid identification of proteins by peptide-mass fingerprinting', **1993**, *Current Biology* **3**:327-332.

Pappin DJ, Hojrup P, Bleasby AJ, ‘Rapid identification of proteins by peptide-mass fingerprinting’, **1993**, *Current Biology* **3**:327-332.

Patterson SD, Spahr CS, Daugas E, Susin SA, Irinopoulou T, Koehler C, Kroemer G, 'Mass spectrometric identification of proteins released from mitochondria undergoing permeability transition', **2000**, *Cell Death and Differentiation* **7**:137-144.

Paul VW, Steinwedel H, 'Ein neues Massenspektrometer ohne Magnetfeld', **1953**, *Zeitschrift fur Naturforschung* **8a**:448-450.

Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R *et al.*, 'A common open representation of mass spectrometry data and its application to proteomics research', **2004**, *Nature Biotechnology* **22**:1459-1466.

Peng J, Schwartz D, Elias JE, Thoreen CC, Cheng D, Marsischky G, Roelofs J, Finley D, Gygi SP, 'A proteomics approach to understanding protein ubiquitination', **2003**, *Nature Biotechnology* **21**:921–926.

Prince JT, Carlson MW, Wang R, Lu P, Marcotte EM, 'The need for a public proteomics repository', **2004**, *Nature Biotechnology* **22**:471-472.

Rauch A, Bellew M, Eng J, Fitzgibbon M, Holzman T, Hussey P, Igra M, Maclean B, Lin CW, Detter A *et al.*, 'Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments', **2006**, *Journal of Proteome Research* **5**:112-121.

Roepstorff P, Fohlmann J, 'Proposal for a common nomenclature for sequence ions in mass spectra of peptides', **1984**, *Biomedical Mass Spectrometry* **11**:601.

Rohlff C, 'New approaches towards integrated proteomic databases and depositories', **2004**, *Expert Reviews in Proteomics* **1**:267-274.

Sadygov RG, Cociorva D, Yates JR 3rd, 'Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book', **2004**, *Nature Methods* **1**:195-202.

Salmi J, Moulder R, Filén JJ, Nevalainen OS, Nyman TA, Lahesmaa R, Aittokallio T, 'Quality classification of tandem mass spectrometry data', **2006**, *Bioinformatics* **22**:400-406.

Santoni V, Molloy M, Rabilloud T, 'Membrane proteins and proteomics: un amour impossible?', **2000**, *Electrophoresis* **21**:1054-1070.

Spahr CS, Susin SA, Bures EJ, Robinson JH, Davis MT, McGinley MD, Kroemer G, Patterson SD, 'Simplification of complex peptide mixtures for proteomic analysis: reversible biotinylation of cysteinyl peptides', **2000**, *Electrophoresis* **21**:1635-1650.

Staes A, Demol H, Van Damme J, Martens L, Vandekerckhove J, Gevaert K, 'Global differential non-gel proteomics by quantitative and stable labeling of tryptic peptides with oxygen-18', **2004**, *Journal of Proteome Research* **3**:786-791.

States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM, 'Deriving high confidence protein identifications from a HUPO collaborative study of human serum and plasma', **2006**, *Nature Biotechnology* **24**:333-338.

Stephan C, Hamacher M, Bluggel M, Korting G, Chamrad D, Scheer C, Marcus K, Reidegeld KA, Lohaus C, Schafer H *et al.*, '5th HUPO BPP Bioinformatics Meeting at the European Bioinformatics Institute in Hinxton, UK--Setting the analysis frame', **2005**, *Proteomics* **5**:3560-3562.

Switzer RC 3rd, Merril CR, Shifrin S, 'A highly sensitive silver stain for detecting proteins and peptides in polyacrylamide gels', **1979**, *Analytical Biochemistry* **15**:231-237.

Tabb DL, McDonald WH, Yates JR 3rd, 'DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics', **2002**, *Journal of Proteome Research* **1**:21-26.

Tabb DL, Saraf A, Yates JR 3rd, 'GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model', **2003**, *Analytical Chemistry* **75**:6415-6421.

Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T, 'Protein and Polymer Analyses up to m/z 100 000 by Laser Ionization Time-of flight Mass Spectrometry', **1988**, *Rapid Communications in Mass Spectrometry* **2**:151-153.

Taylor JA, Johnson RS, 'Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry', **2000**, *Analytical Chemistry* **73**:2594-2604.

Van Damme P, Martens L, Van Damme J, Hugelier K, Staes A, Vandekerckhove J, Gevaert K, 'Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis', **2005**, *Nature Methods* **2**:771-777.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al.*, 'The sequence of the human genome', **2001**, *Science* **291**:1304-1351.

Wang H, Hanash S, 'Intact-protein based sample preparation strategies for proteome analysis in combination with mass spectrometry', **2005**, *Mass Spectrometry Reviews* **24**:413-426.

Wang S, Regnier FE, 'Proteomics based on selecting and quantifying cysteine containing peptides by covalent chromatography', **2001**, *Journal of Chromatography A* **924**:345-357.

Wang S, Zhang X, Regnier FE, 'Quantitative proteomics strategy involving the selection of peptides containing both cysteine and histidine from tryptic digests of cell lysates', **2002**, *Journal of Chromatography A* **949**:153–162.

Washburn MP, Wolters D, Yates JR 3rd, 'Large-scale analysis of the yeast proteome by multidimensional protein identification technology', **2001**, *Nature Biotechnology* **19**:242–247

Washburn MP, Ulaszek R, Deciu C, Schieltz DM, Yates JR 3rd, 'Analysis of quantitative proteomic data generated via multidimensional protein identification technology' **2002**, *Analytical Chemistry*, **74**:1650-1657.

Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, Duncan MW, Harris R, Williams KL, Humphery-Smith I, 'Progress with gene product mapping of the Mollicutes', **1995**, *Electrophoresis* **16**:1090-1094.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al.,* 'Initial sequencing and comparative analysis of the mouse genome', **2002**, *Nature* **420**:520-562.

Wilkins MR, Appel RD, Van Eyk JE, Chung MC, Gorg A, Hecker M, Huber LA, Langen H, Link AJ, Paik YK *et al.*, 'Guidelines for the next 10 years of proteomics', **2006**, *Proteomics* **6**:4-8.

Wolters DA, Washburn MP, Yates JR 3rd, 'An automated multidimensional protein identification technology for shotgun proteomics', **2001**, *Analytical Chemistry* **73**:5683–5690.

Wong C, So MP, Dominic Chan TW, 'Origins of the proton in the generation of protonated polymers and peptides in matrix-assisted laser desorption/ionization', **1998**, *European Journal of Mass Spectrometry* **4**:232-233.

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B, 'The Universal Protein Resource (UniProt): an expanding universe of protein information', **2006**, *Nucleic Acids Research* **34**(Database issue):D187-D191.

Yates JR 3rd, Speicher S, Griffin PR, Hunkapiller T, 'Peptide mass maps: a highly informative approach to protein identification', **1993**, *Analytical Biochemistry*, **214**:397-408.

Zenobi R, Knochenmuss R, 'Ion formation in MALDI mass spectrometry', **1998**, *Mass Spectrometry Reviews* **17**:337-366.

Zhang H, Li XJ, Martin DB, Aebersold R, 'Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling, and mass spectrometry', **2003**, *Nature Biotechnology* **21**:660–666.

Zhang Z, 'De novo peptide sequencing based on a divide-and-conquer algorithm and peptide tandem spectrum simulation', **2004**, *Analytical Chemistry* **76**:6374-6383.

Zhao S, Zhong F, Chang JH, Zhu Z, 'Studies of some characters of matrix influencing desorption ionization processes in matrix-assisted lased desorption ionization mass spectrometry', **1997**, *Analytical Science* **13**:45-48.