Dissertation zur Erlangung des Doktorgrades
Der Fakultät für Chemie und Pharmazie
Der Ludwig-Maximilians-Universität München

High accuracy mass spectrometric peptide identification as a discovery tool in proteomics

Alexandre Zougman

Aus

Odessa, Ukraine

2009

Erklärung

Diese Dissertation wurde im Sinne von § Abs. 3 der Promotionsordnung vom 29. Januar 1998 von Herrn Prof. Dr. Matthias Mann betreut.

Ehrenwörtliche Versicherung

Diese Dissertation wurde selbständig, ohne unerlaubte Hilfe erarbeitet.

München, am February 19, 2009

.....
(Unterschrift des Autors)

Dissertation eingereicht am February 19, 2009

1. Gutachter Prof. Dr. Matthias Mann
2. Gutachter Prof. Dr. Jacek R. Wiśniewski

Mündliche Prüfung am March 23, 2009

Contents

Abbreviations

amu: atomic mass unit
CID: collision induced dissociation
CNS: central nervous system
CSF: cerebrospinal fluid
EST: expressed sequence tags
ESI: electrospray ionization
ET: extended protein
FT: Fourier transform
FT-ICR: Fourier transform ion cyclotron resonance
FWHM: full width of the peak at half of its maximum height
GC: gas chromatography
HCD: high energy C-trap dissociation
HPLC: high performance liquid chromatography
ICR: ion cyclotron resonance
LC-MS: liquid chromatography mass spectrometry
LTQ: linear trap quadrupole
m/z: mass-over-charge ratio
MALDI: matrix-assisted laser desorption ionisation
MS: mass spectrometry
MS/MS: tandem mass spectrometry fragmentation
NanoLC: nanoflow liquid chromatography
NanoES: nanoelectrospray
ppm: part per million
PTM: post-translational modification
Q-TOF: quadrupole time of flight
RF: radio frequency
RP: reverse phase
RT: retention time
TOF: time of flight
UV: ultraviolet light wavelength

Summary

The recent achievements of proteomics are mainly due to the developments in the high accuracy and resolution mass spectrometry. Protein identification by conventional proteomics "shotgun" approach involves enzymatic cleavage and consequent identification of the resulting peptide products. Typically, several peptides belonging to the same unique protein are produced by this process and, considering the improved sensitivity and separation abilities of the contemporary LC-MS instrumentation, it is feasible to identify a known protein with at least two unique peptides by submitting a query to a protein database. In order to be time efficient, a typical LC-MS run includes high accuracy identification of the precursor peptide mass during the MS scan event and significantly less accurate but fast MS/MS peptide fragmentation signature profiling. This scheme generally satisfies very stringent protein identification requirements. However, identification of endogenous peptides such as neuropeptides, for example, often must rely on one peptide only and, hence, has to be much more accurate in order to provide the needed level of confidence. Likewise, the MS/MS data have to be very accurate for identification and characterization of novel post-translational modifications. Furthermore, *de-novo* sequencing of unknown proteins not present in protein databases frequently relies initially on the discovery of only one peptide which provides the basis for the following characterization of the unknown protein. For the above-mentioned cases the availability of high accuracy data in both MS and MS/MS modes is of the utmost importance. In my work I utilized high accuracy mass spectrometry for discovery of novel extended forms of nuclear proteins and underlying mRNA editing events, and, also, for characterization of the human cerebrospinal fluid peptidome and proteome.

"Tout progrès scientific est un progrès de method."

"All scientific progress is progress of a method."

René Descartes

Introduction

The extraordinary achievements of current proteomics are based largely on successful developments in the fields of mass spectrometry (MS) and separation science. Once joined, the two disciplines provided a powerful tool to investigate the protein universe.

Mass spectrometry

Mass spectrometry is the science of determining mass-to-charge ratios (m/z) of ionized molecules. In 1897 Sir Joseph John Thomson, a British physicist, in a series of experiments with cathode rays discovered the existence of the electron and determined its mass to charge ratio m/z (*1*). He was awarded the Nobel Prize in physics in 1906 and is considered a founder of the field of mass spectrometry. It took more than half a century of research and development before a first commercial mass spectrometer appeared, and almost a century to fully (or *almost* fully) seize the analytical potential of the MS instruments. This development has gone in many directions – with different instrument designs using different physical principles. However, the essence of any mass spectrometer is the same – to determine the m/z of an introduced ion. William Stephens of the University of Pennsylvania presented the idea of the time-of-flight (TOF) mass analyzer at the 1946 American Physical Society Meeting in Cambridge. His concept was to accelerate all ions to the same kinetic energy. Thus an accelerated ion would have the same kinetic energy as any other ion (provided they had the same charge), with the velocity depending solely on m/z (*2*). Wiley and McLaren of Bendix Corporation in Detroit introduced a TOF focusing scheme that improved mass resolution by correcting energy distributions of the ions (*3*). As the result, Bendix was the first to commercialize TOF instruments in the 1950s. In 1973 Boris Mamyrin, a

Russian scientist who worked at the Leningrad Physico-Technical Institute, described the reflectron and the energy compensating mirror system which corrected for the effects of the kinetic energy distribution of the ions in the TOF tube. This invention resulted in a dramatic increase of TOF resolution (*4*). Needless to say, the 1970s were also the years when the affordable fast timing electronics became available which was crucial for the further development of the time-of-flight field in order to handle the huge data flow. In 1953 German physicists Wolfgang Paul who worked at Bonn University and Helmut Steinwedel at University of Wurzburg registered a patent describing the quadrupole mass analyzer (*5*). They showed that the static and radio frequency (RF) quadrupole oscillating electric fields can be used to act as an m/z separator. This discovery led to the development and commercialization of triple quadrupole mass analyzers and 3D ion trapping instruments. The first three-dimensional versions of the device now known as a Paul trap (or *ionenkäfig*, the "ion cage", as Paul liked to call it) were developed by Wolfgang Paul and Hans Dehmelt (University of Washington) and allowed to detect and measure the ions while stored. Later developments added the ability of mass-selective storage and ejection to the ion-trap resulting in the production of a reliable commercial instrument by George Stafford at the Finnigan company (*6*). For the "development of the ion trap technique" Paul and Dehmelt were awarded Nobel Prize in Physics in 1989. In 1932 Ernest Lawrence described a cyclotron principle and its application to high energy physics which brought him a Nobel Prize in Physics in 1939 (*7*). More than forty years afterwards, in 1974, Fourier Transform Ion Cyclotron Resonance (FT-ICR) mass spectrometry was introduced by Alan Marshall and Mel Comisarow (*8*). In this method the ions present in the cyclotron cell are excited by broadband RF. In a strong fixed magnetic field each ion of a given mass will have its characteristic cyclotron frequency. The circulating tight packets of such ions induce image currents which are decoded by Fourier Transform (FT) analysis. Even though their maintenance is cumbersome because of the need for constant supply of liquid nitrogen and helium to keep the superconductive magnets going, FT-ICR mass spectrometers are the most accurate mass measuring instruments. Different mass analyzers can be coupled together in order to use their properties to the fullest potential – current mass spectrometry is dominated by such *hybrid* instruments. The successful hybrid quadrupole-TOF (Q-TOF) instrument, for example, was developed in the late 1990s at MDS SCIEX, Concord, Canada and combined the guiding and isolating properties of a quadrupole mass filter with the high accuracy of the TOF detection

analyzer (*9*). It took more than 80 years since the original discovery of orbital trapping by Kingdon in 1923 (*10*) before the commercial hybrid linear triple quadrupole (LTQ)-Orbitrap instrument by Thermo Finnigan hit the market. This hybrid instrument combines the guiding, trapping and detection properties of the linear quadrupole trap with the high accuracy m/z detection ability of the Orbitrap - the work by the Russian physicist Alexander Makarov at HD Technologies Inc., Manchester, UK built the foundation for the successful entry of this machine into the proteomics world. Using Kingdon's concept he created a high accuracy and high resolution instrument, the orbitrap, in which the ions could be trapped in the absence of any magnetic or radio frequency field, and ion stability is achieved only due to ions orbiting around an axial electrode. The oscillations of the ions are revealed using image current detection and are decoded by Fourier Transformation (*11*). The accuracy and resolving power of the orbitrap is comparable with that of FT-ICR instruments but without the need for expensive maintenance.

Generating ions

Availability of instruments that allowed mass determination of molecules was attractive for biologists. However, in order to determine the m/z of a molecule this molecule has to be ionized and introduced into the gas phase. This requirement was less of a challenge for small organic molecules, which were forced into the gas phase by heating leading to vaporization and subsequently converted to ions by electron or chemical ionization (*12, 13*). Both methods resulted in post-source fragmentation of the generated ions, with the lower energy chemical ionization being somewhat less destructive of the generated ions. For obvious reasons, proteins and polypeptides could not be introduced into the gas-phase by such a harsh treatment. The question of finding the proper ionization method was of the utmost importance for biological mass spectrometrists. The solutions appeared with the publications of John Fenn's group at Yale University, USA (*14, 15*) which ignited the development of the electrospray ionization (ESI) field, and Michael Karas and Franz Hillenkamp at Frankfurt University, Germany (*16*), and Koichi Tanaka at Shimadzu Inc., Japan (*17*) which resulted in the development of the matrix-assisted laser desorption ionization (MALDI) methods. TMALDI and ESI provided "soft", non-destructive ways of ionization for biomolecules and, eventually, lead to the establishment of

mass spectrometry-based proteomics. These achievements were recognized by the awarding of the 2002 Nobel Prize in Chemistry to John Fenn and Koichi Tanaka.

MALDI - matrix-assisted laser desorption ionisation

In MALDI the analyte is mixed with an excess of matrix molecules. Upon drying, the analyte molecules are incorporated into the matrix crystals. The matrix absorbs energy at the wavelength of a laser and transfers it into the analyte. This causes the analyte to vaporize and form ions (Figure 1). MALDI ion sources are typically combined with TOF mass spectrometers. In 1989 Ronald Beavis and Brian Chait reported the discovery of suitable UV matrices for polypeptide characterization by MALDI (*18*). The UV-excitable matrices were based on cinnamic acid and, mostly, have not changed ever since. Reliable UV lasers for MALDI became available in the late 1980s. These developments favored general acceptance of MALDI-TOF instrumentation in biological research. The MALDI process produces predominantly singly-charged (1+) ions which are not fragmented easily by the low energy collision induced dissociation (CID) process employed in many contemporary mass spectrometers. If one has to only determine a mass of a protein, having a 1+ ion is of advantage because no peak deconvolution post-processing is involved – in many cases deconvolution of the multiply charged envelopes of large proteins is a very complicated, if not close to impossible, process. Furthermore, a pure protein entity or simple protein mixture can be identified after digestion with a specific protease of choice. The resultant peptides are analyzed by a MALDI-TOF instrument and their corresponding masses compared to a database containing information about the peptide masses calculated for each protein sequence. This identification method is called “peptide mass fingerprinting”.

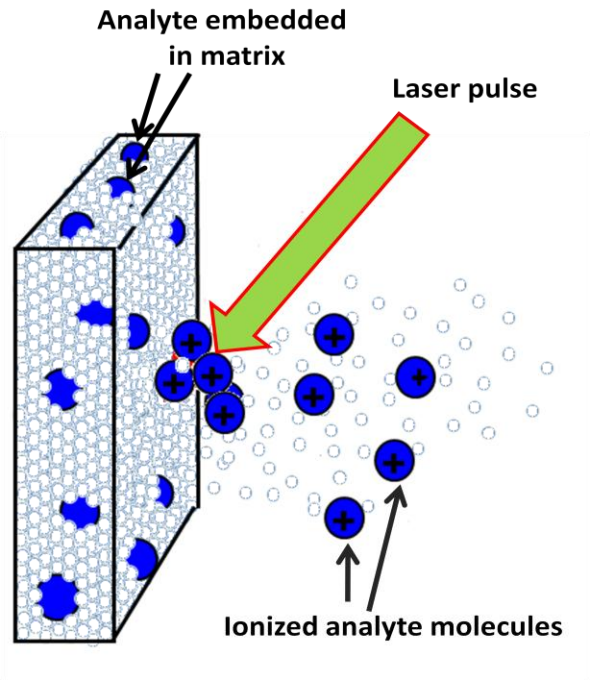


Figure 1. MALDI (matrix-assisted laser desorption ionization) process – analyte is vaporized together with matrix molecules forming predominantly 1+ ions.

However, if one needs to analyze a complex protein mixture, as is usually the case in biology, or the need arises to further characterize the ionized polypeptide by finding out its detailed primary structure, the peptide ions must be isolated and fragmented. In such a case, multiply charged ions are advantageous.

ESI – electrospray ionization

The ESI process transforms solubilized analyte molecules into gaseous ions at atmospheric pressure. The electric field is applied between the spraying capillary containing the analyte and the entry point of a mass spectrometer. Ions accumulate at the liquid surface of the capillary through which liquid is passed (Figure 2). When the surface tension in the cone is exceeded, a conical shape known as ‘Taylor cone’ is formed charged micro-droplets with a diameter less than or equal 1-10 μm are produced. The solvent quickly evaporates after the droplets are formed,

electric surface charge density of the droplet increases and, finally, the droplet bursts releasing the ionized analyte species. ESI allows formation of multiply-charged ions, which makes fragmentation of the ions significantly easier compared with the 1+ species. As a result, if one needs to find an internal signature of an ion (ionized peptide) one isolates the ion, fragments it and performs interpretation of the resulting fragmentation pattern. ESI is easily interfaced with liquid chromatography-based instrumentation therefore allowing on-line separation of complex samples. Due to the multiple-charged nature of the ions generated by ESMS, their m/z ratios are "collapsed" relative to their nominal masses allowing the use of simple (with narrow *m*/*z* range of detection) and relatively inexpensive quadrupole instruments.

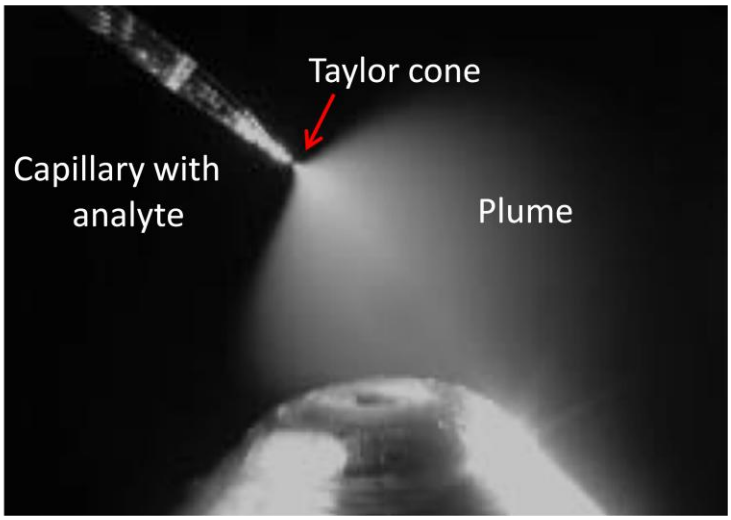


Figure 2. Electrospray ionization process – analyte is transformed into a mist of charged droplets.

Assessing the limits of protein detection

The early 1990s was the time when biological mass spectrometrists started to test the limits of protein detection. Mark Emmett and Richard Caprioli demonstrated low attomole detection limits with electrospray (*19*). They reported the development of the micro-electrospray ion source which significantly increased the sensitivity of peptide identification by ESI. The ion source was modified to accommodate a capillary needle. The needle was coated with C_{18} chromatographic packing material and the liquid flow rate was in the range of 500 nl/min. The authors noted the capability of their spraying device to be positioned closer to the MS analyzer entry point, and suggested that the decreased distance from the mass spectrometer nozzle and low flow rate gives a spray pattern of narrow dispersion which allows more analyte to be drawn into the analyzer leading to better sensitivity. In 1994 Ole Vorm and Matthias Mann demonstrated the attomole detection capabilities of MALDI (*20*). They described a sample preparation procedure in which matrix and sample handling are completely decoupled - first, a drop of matrix is deposited onto the MALDI plate, second, a drop of analyte solution is deposited onto the layer of the matrix microcrystalls. The simple sample preparation procedure leads to improvement in sensitivity, with peptides routinely analyzed at the attomole range. The same year, Matthias Wilm and Matthias Mann reported development of the nanoelectrospray ion source (nanoES). The nanoES source employed pulled capillaries with 1-2 μm spraying orifice and about 20 nl/min flow rate which lead to very small droplet sizes and improved desolvation efficiency. The sample is loaded directly into the spraying capillary. The nanoES flow is not forced by solvent pumps – it is driven by the electrospray process itself (*21*). In 1996 Andrej Shevchenko and Matthias Mann established the feasibility of the microcharacterization of silver-stained proteins by MALDI and NanoES tandem mass spectrometry (*22*). They proved the compatibility of the highly sensitive silver-staining method with microanalytical protein characterization. The sample detection limit in this case was 10-100 times lower than that obtained with Coomassie staining. They also pointed out the advantage of NanoES MS/MS profiling over MALDI peptide mass fingerprinting when identifying less abundant protein components of complex protein mixtures – indeed, the identification of the low level proteins is often based only on few peptides which are “hiding” in the background of the peptide products of more abundant proteins.

Isotopic envelopes

Most chemical elements exist as diverse masses in nature, the isotopes of the element. Proteins are composed of carbon (C), hydrogen (H), nitrogen (N), oxygen (O), sulfur (S) and phosphorus (P). The natural abundances of the corresponding isotopes are shown in Figure 3.

Element	Mass (Da)/Abundance
C12	12.0000/ 98.89%
C13	13.0034/ 1.11%
H1	1.0078/99.985%
H2	2.0141/ 0.015%
N14	14.0031/99.634%
N15	15.0001/ 0.366%
O16	15.9949/99.762%
O17	16.9949/ 0.038%
O18	17.9992/ 0.200%
S32	31.9721/95.02%
S33	32.9715/ 0.75%
S34	33.9679/ 4.21%
S36	35.9671/ 0.02%
P31	30.9738/100.00%

Figure 3. Isotopic abundances of protein elemental components.

About 1.11 % of the naturally occurring carbon ^{13}C atoms are 1.0034 Da heavier than the C_{12} atoms. This means that from a random set of 100 carbon atoms about 99% will have a mass of 12.0000 Da and 1% will have a mass of 13.0034 Da. The larger the polypeptide mass, the higher the probability of C_{13} incorporation into this molecule and the more prominent the isotopic effect

(Figure 4). In fact, large polypeptides show inversion of the isotopic envelope when the first C_{12}-only *monoisotopic* peak is less intense than the adjacent C_{13}-containing peaks (Figure 4C). Because it is a major building block of polypeptides and has distinct isotopic diversity, carbon has the most prominent effect on the polypeptide isotopic profile.

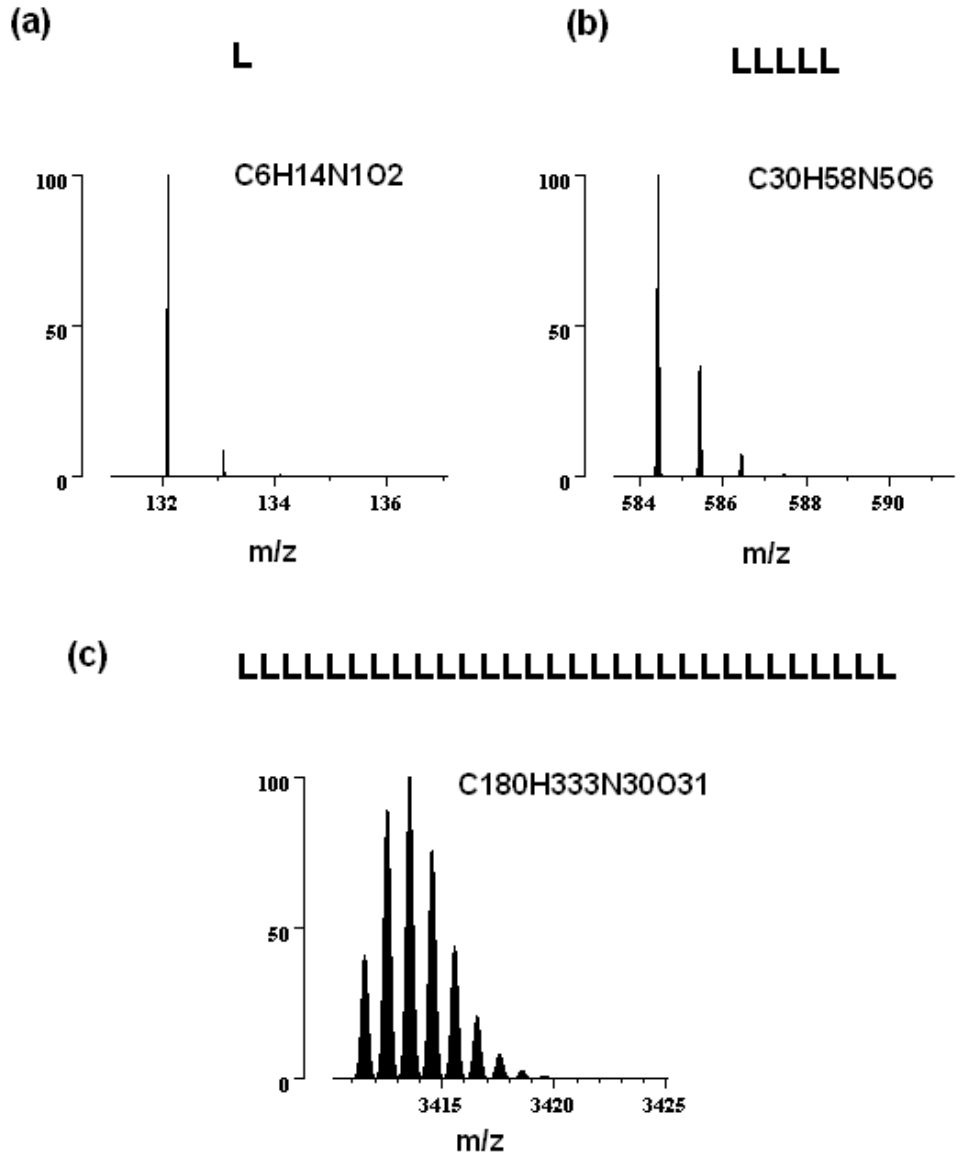


Figure 4. Isotopic envelopes of singly-charged leucine (L)-containing peptides.

Accuracy and resolution

The *mass accuracy* of a mass spectrometer is defined as the difference between the measured mass and its calculated value. Traditionally *relative mass accuracy* is reported and defined in *parts per million* (ppm) – the *mass accuracy* divided by a calculated value of the measured mass. High mass accuracy is directly related to the ability of a mass spectrometer to separate the adjacent peaks, to its *resolution*. The resolution can be calculated by dividing the m/z of a peak by its width at a certain height (m/z divided by Δm/z). A common definition of resolution is *Full Width of the peak at Half of its Maximum height*, abbreviated as 'FWHM', the Δm/z value in this case is taken at 50% of the peak height (Figure 5). The difference between the resolution of different instruments is enormous. An average quadrupole type instrument has a FWHM resolution of 2000 in the 400-800 m/z range (it is still common to use the "unit resolution" definition for the triple quadupole analyzers which refers to the ability of the instrument to separate each mass from the next integer mass), whereas advanced hybrid instruments offer resolution of up to 100 000 in this range and even more. For the purpose of identification the accuracy of a mass spectrometer is of the utmost importance – indeed, the more accurately one determines the masses of peptides and their fragments the better the probability of correct identification. The linear trap quadrupole instrument (LTQ), for example, has an accuracy of about 150 ppm at 700.0 m/z; the Quadrupole-TOF analyzer (Q-TOF) offers about 20-30 ppm accuracy at this m/z value, and the LTQ-Orbitrap achieves 0.5-1 ppm.

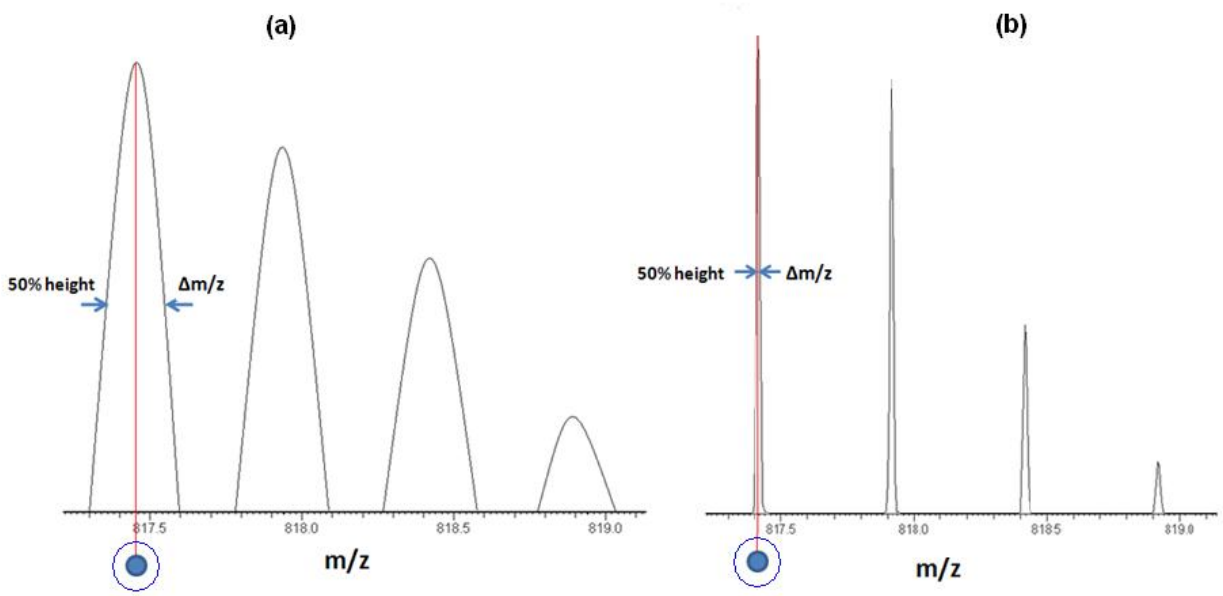


Figure 5. Full Width of the peak at Half of its Maximum height (FWHM) resolution of (a) 4000 and (b) 50000.

Analyzing the ionized peptides

It is important to stress that a mass spectrometer determines the mass-over-charge ratio and not the mass of the introduced ions. Most common proteomics experiments utilize positive ionization mode for peptides. The resulting charge of the peptide ion directly corresponds to the number of accepted protons (1+ means 1H+ accepted; 2+ means 2H+ accepted etc.). The isotopic nature of the ionized protonated peptides allows elucidating of their charge and, consequently, mass (Figure 6). For example, the mass difference of 0.5017 (1.0034/2) atomic units (amu) between the peaks of the isotopic envelope corresponds to the charge of 2. Multiplying the m/z value of the first, monoisotopic, peak of the envelope by its charge and

subtracting the mass of the corresponding accepted protons (1.0073 Da per 1H+) yields the experimental monoisotopic mass of the peptide.

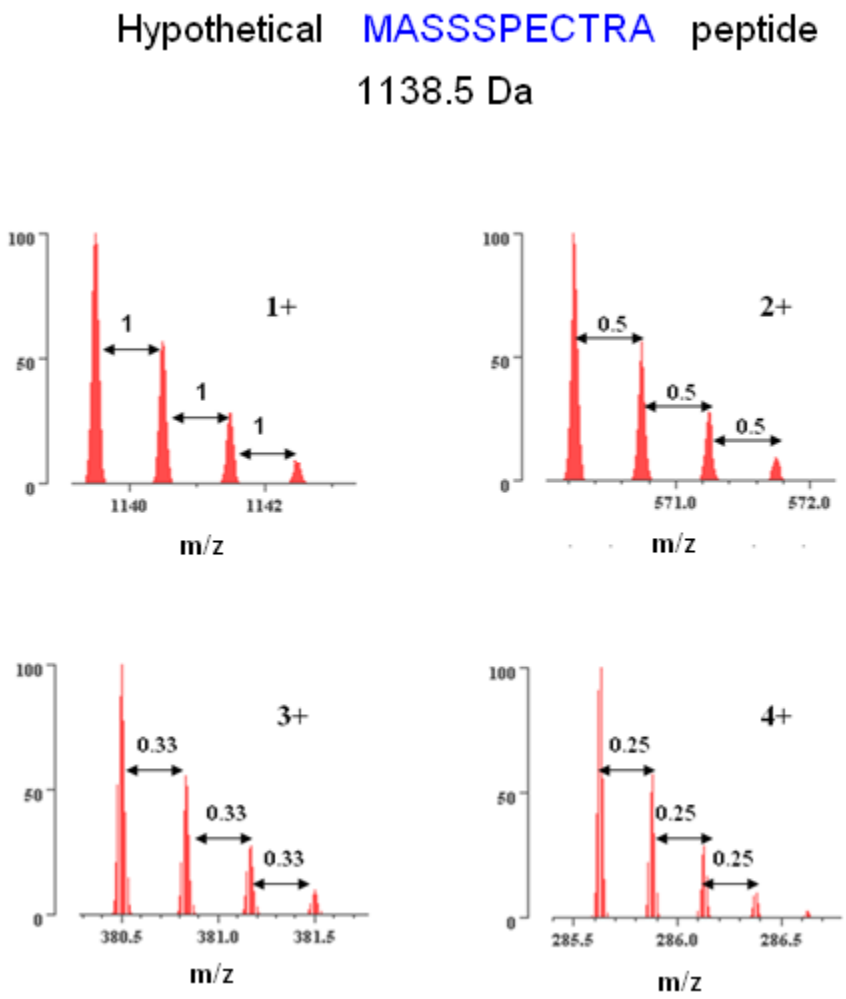


Figure 6. Charge profiles of the hypothetical *MASSSPECTRA* peptide.

Tryptic peptides and fragmentation

In a typical proteomics experiment a protein mixture is digested by an enzyme thus creating a mixture of peptides. The most common protease is trypsin which cuts protein C-terminally at arginine (R) or lysine (K) residues and gives rise to peptides carrying at least one basic residue (except the C-terminal peptide of the protein). At pH 3, Asp- and Glu- are uncharged and any proteolytic peptide has the net charge of at least 2+ where one charge is located C-terminally at the basic residue and another charge is localized at the N-terminus.

Figure 7. Typical tryptic peptide.

The opposite location of the charges facilitates peptide fragmentation at the amide bond in the most commonly used low energy collision induced dissociation (CID) process creating *y* and *b* ion series according to Roepstorff-Fohlmann-Biemann nomenclature (*23*).

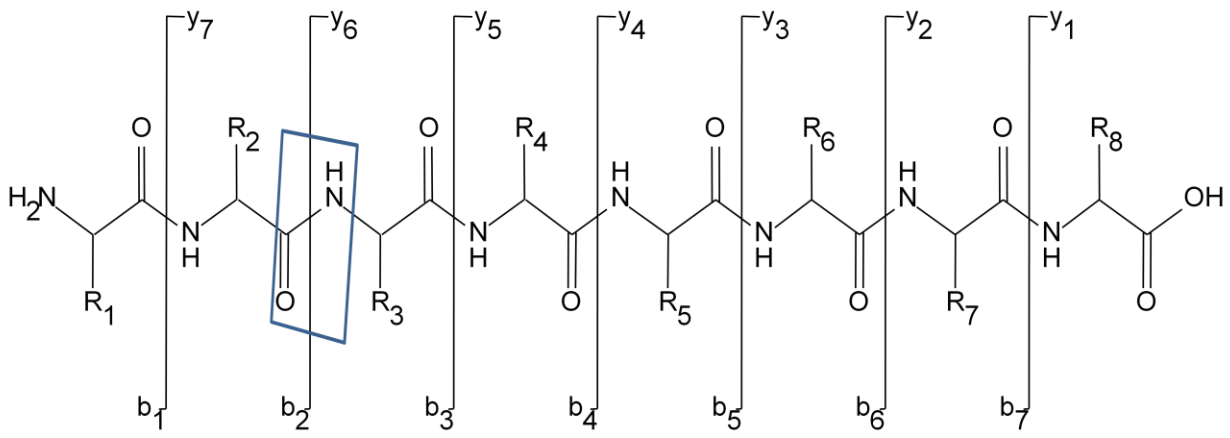


Figure 8. Roepstorff-Fohlmann-Biemann nomenclature for low energy CID. The blue shape indicates the peptide bond that is fragmented, in this case giving rise to the b_2 and y_6 fragment ions.

This convenient placement of the charges is typical of tryptic peptides generated by enzymatic digestion but it is not usual for endogenous peptides. Their amino acid sequence does not follow strict rules - the basic residues, if any, are distributed over the sequence which makes the analysis of such molecules much more challenging than that of tryptic peptides. In addition, functional endogenous peptides such as neuropeptides are often post-translationally modified. Modifications such as phosphorylation or glycosylation are very labile and their loss is often the predominant fragmentation channel, reducing CID fragmentation efficiency.

NanoLC-MS

Currently capillary columns are the most effective chromatographic separation devices available. The capillary column format results in a dramatic increase of the LC-MS sensitivity due to the concentration effect (Figure 9). Capillary columns were originally introduced for gas chromatography (GC) by Marcel Golay of Perkin Elmer at the *1958 Symposium on Gas Chromatography* (*24*) more than fifty years ago. He presented the separation of C8 hydrocarbons and the xylene isomers on a 50 m x 0.25 mm stainless steel diisodecyl phthalate coated column together with about 90 equations supporting his elegant concept. About a decade afterwards, the

idea of using capillary separations for liquid chromatography (LC) was clarified in publications of Horvath *et al.* (*25*) and Ishii (*26*). The field evolved in parallel with research and development in the area of the chromatographic stationary supports. The requirements of pharmaceutical chemistry have been the major driving force for the advancement of chromatographic techniques because the quality of pharmaceutical products is determined by the quality of their chromatographic analysis. The expansion of chemically bonded hydrophobic stationary phases which were named *reverse phase* (RP) packings in High Performance Liquid Chromatography (HPLC) began in 1970 when Jack Kirkland of Du Pont de Nemours & Co, Delaware synthesized a hydrophobic bonded phase by coupling a poly-*n*-octadecylsiloxane to the surface of a pellicular matrix (*27*). Soon after, Ron Majors prepared bonded phases by the reaction of organotrichlorosilanes with microparticulate silicas (*28*). The challenge was to synthesize RP silicas to separate basic analytes with sufficiently symmetrical peak shape and reproducible retention coefficients. The problem was solved by manufacturing silicas with a reduced acidity and a high purity and appropriate *n*-alkyl functionality. Furthermore, many separations, especially peptide separations, were performed using acidic mobile phases and organic mobile phases at pH 2–3. This required RP packings with a high stability at acidic pH. Jack Kirkland and his team at Rockland Technologies, Delaware successfully developed a number of surface chemistries, including the introduction of steric protection of the siloxane bond that holds the bonded phase to the silica surface and bidentate bonded phases with two anchoring siloxane bonds (*StableBond* and *Extend*, *re*spectively). The introduction of spherical particles improved packing stability of the columns. Further reduction of the diameter of the stationary phase particles from initially 10 – 5 µm in 1975, to 3 µm in 1978, and 1.5 µm in 1990 dramatically improved HPLC performance. The evolution of capillary columns culminated in the creation of fused silica columns and was directly connected to the developments in the fiber optics industry. In 1979 at the *Third Hindelang Symposium* R.D. Dandeneau and E.H. Zerenner reported (*29*) on production and use of flexible fused-silica columns. Their column tubing was adapted from the fiber optics tubing already manufactured at Hewlett-Packard, California. To prevent the cracks and breaks of the fragile fused silica, they coated the outside of the tubing immediately after drawing with silicone rubber (nowadays the fused silica tubing is coated with polyimide). Microscale LC-MS/MS was first used to analyze peptides by Donald Hunt (*30*). A typical contemporary nanoLC capillary column has a length of 10-15 cm; inner diameter 75-100 µm and

is packed with 3 or 5 µm C_{18} reversed phase particles. Most of the commercially available capillary columns are provided in the fritted format leaving it up to the end-user to choose spray capillaries with opening of 5-8 µm which is attached to the column through the tee union (Figure 10). Such a set-up though acceptable to many has substantial drawbacks - in addition to introducing additional "void volume" leading to peak broadening and post-column mixing, the spray capillary opening is easily clogged by stray pieces of packing material. Some groups suggested to pack media particles into the tapered capillary columns with the particles diameter being larger than the column opening (*31*). In this case, the tapered end acts both as emitter and restrictor for the packed material. The reported setup results in significant increase in backpressure of a chromatography system (typically more than 250 bars), common fluidic blockage and deleterious consequences on the pump performance. In 2002 Yasushi Ishihama and Matthias Mann described a so-called "arch" concept of the capillary column packing (*32*). They suggested packing particles of a *lesser* diameter than a tapered end of the capillary column which also serves as an emitter. The particles produce self-assembled frits based on the stone bridge arch principle (Figure 11, 12). The created self-assembled frit is stable, results in a backpressure less than 180 bar for a typical capillary nanoflow setup, thus reducing the backpressures used in capillary LC (up to 300 bars) and eliminating post-column dilution of the analyte. The setup combining nano-LC system and spraying capillary column is nowadays the standard for introducing peptide samples into a mass spectrometer (Figure 13).

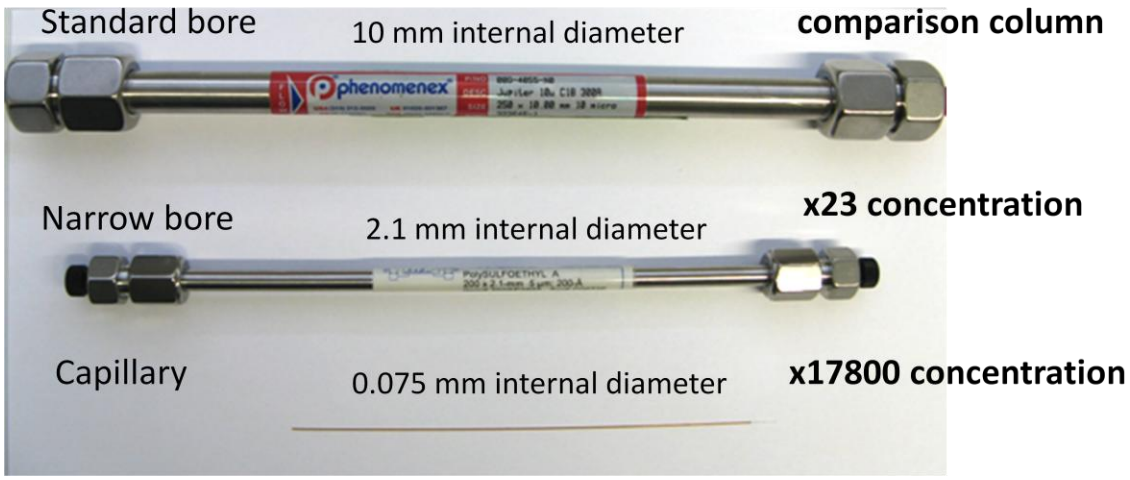


Figure 9. Capillary columns with inner diameter of 0.075 mm or less dramatically increase LC-MS sensitivity.

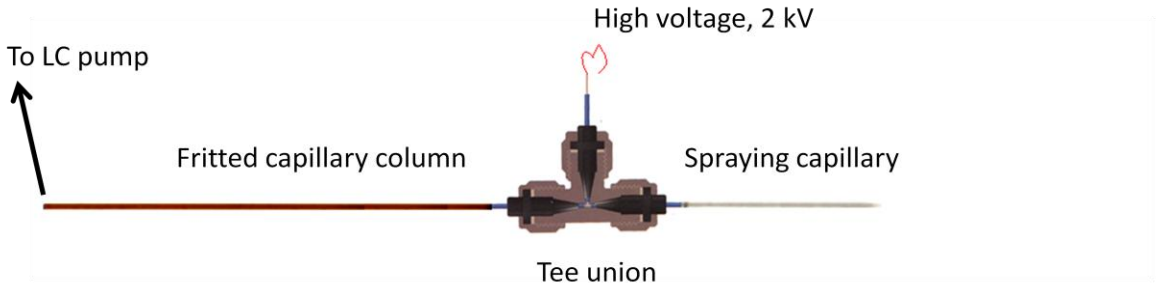


Figure 10. NanoLC connection for a fritted capillary column.

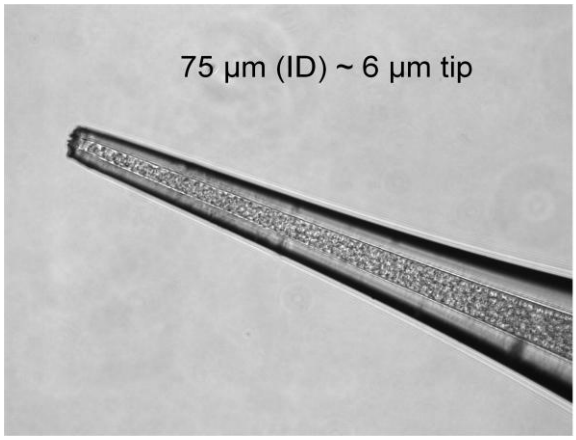


Figure 11. Fused silica spraying capillary packed with 3µm C18 silica particles. X400 magnification.

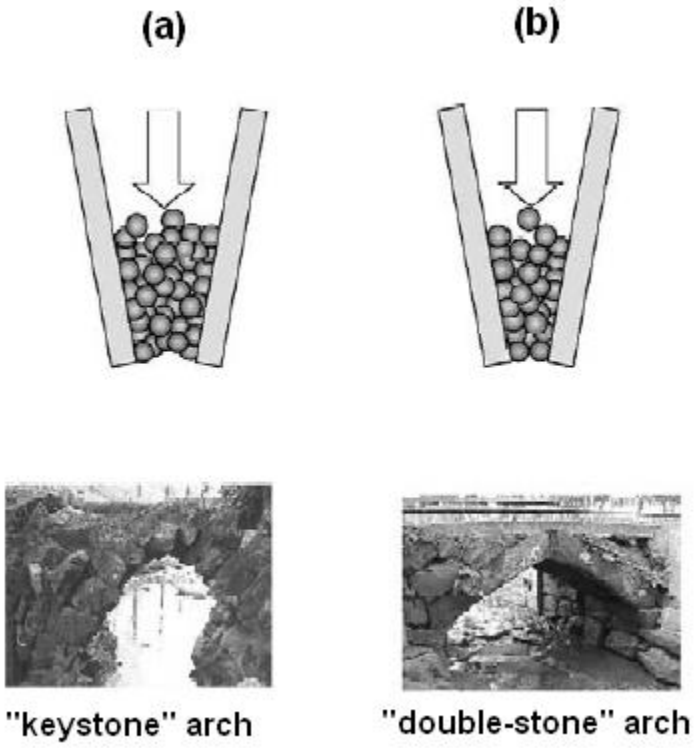


Figure 12. Columns utilizing arch types found in stone-bridges. (a) Tapered column with the ‘‘keystone’’ arch, (b) tapered column with the ‘‘double-stone’’ arch; (left picture) a stone-bridge with the keystone arch, (right picture) a stone-bridge with the double-stone arch. Taken from ref. (*32*).

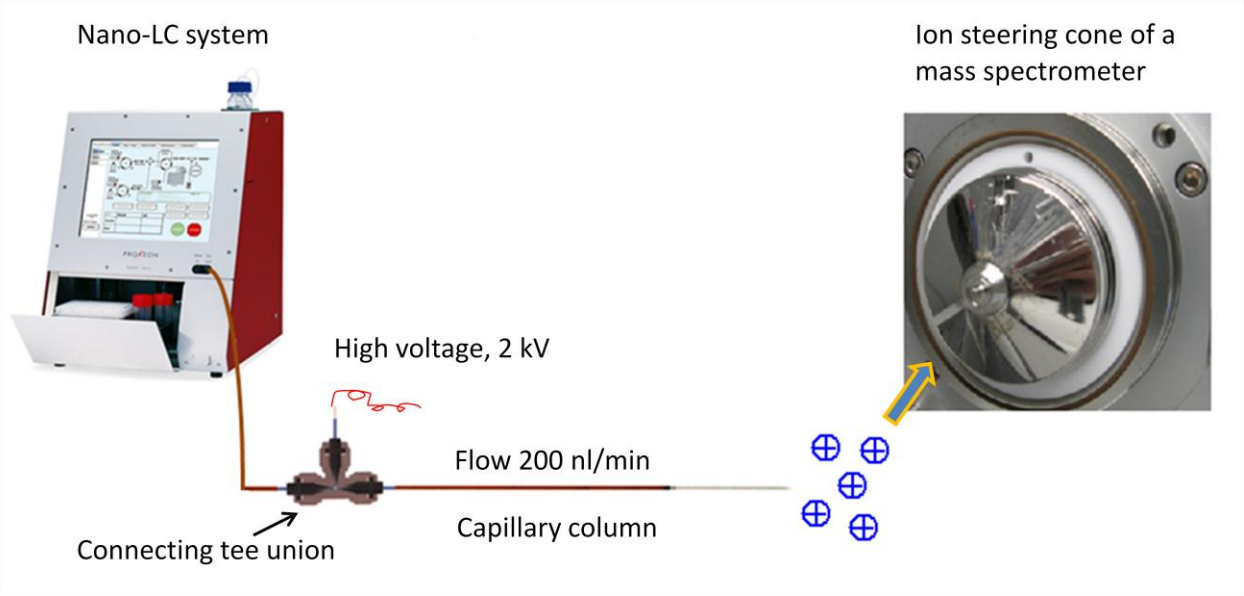


Figure 13. Nano-LC set-up using spraying capillary column.

The LC-MS/MS runs and protein database searches

Modern proteomics profiling methodology are based mostly on the so-called "shotgun" approach when MS data on the peptides resulting from the enzymatic digest of the protein mixture are collected at a high resolution and mass accuracy, and MS/MS fragmentation data are recorded at high speed with low resolution and accuracy maximizing sensitivity and throughput (*33*). The full MS scan provides information to choose a precursor ion mass for the following fragmentation event (Figure 14). The MS scan results in information about the mass and intensity of a precursor peptide (Figure 15a). The subsequent MS/MS fragmentation provides an amino acid sequence signature of the peptide (Figure 15b). Sequence database search algorithms are used to assign sequence information to MS/MS spectra (*34-36*). All peptide candidate sequences that match the experimental peptide mass within the allowed mass deviation are selected from an *in silico* digested protein sequence database. Each candidate is further processed at the MS/MS level by correlating the experimental and theoretical peptide fragmentation

patterns also within the allowed mass deviation. This produces a probabilistic score for the likelihood that the match is a false positive. Any of the 'search engines' utilizes a scoring scheme calculating the significance of peptide assignments and provides theoretically or empirically derived statistical thresholds for assessment of peptide identifications. Figure 16a shows the results of a database search for a particular peptide. All the peptide identifications per unique protein are aggregated and presented as a coverage map (Figure 16b). If the protein exists in the database that is searched and relevant MS and MS/MS spectra of reasonable quality are present, than the identification of such a protein by conventional computational methods is an easy task. Nevertheless, in a typical proteomics experiment the peptide identification rate rarely exceeds 30%. One of the reasons for this are post-translational modifications (PTM) of proteins which result in a peptide mass shift (delta-mass, ΔM). The conventional search engines perform poorly in this case – before submitting the search a user has to have an *a priori* knowledge of the modifications (which is often not the case) and specify all potential modifications as "variable". Allowing for many variable modifications in the database search increases the rate of false positives and requires a much higher score threshold for confident identification of peptides, which leads to an enhanced number of false negative results (*37*). This problem was approached with the recent introduction of the *Modificomb* software tool (*38*), which is based on the concept that most of the PTMs are present in substoichiometric amounts. Thus each peptide "family" will consist of an unmodified base peptide and several modified or 'dependent' peptides with the same sequence but with a PTM. *Modificomb* searches for such families, "combs out" from large data arrays pairs of peptides with sequence identities one of which is a base peptide and the other of which is a dependent peptide (Figure 17). Often, the dependent peptide has similar chromatographic properties and thus appears in a narrow retention time (RT) window around the base peptide. *ModifiComb* separates the function of peptide identification from the function of the PTM assignment and is even able to use a lower information content from MS/MS data of modified peptides than is required for reliable database identification of unmodified peptides by conventional search engines. However, if the protein is absent in the database its identification is not possible using sequence database search algorithms – in such a case *de-novo* sequencing is needed. At this moment, several commercially available software tools for automated *de-novo* sequencing of the peptide spectra exist (*39-41*). In my opinion, even though these tools are extremely helpful in guiding users towards interesting hits, they are still not reliable and in many

cases it is the expert judgment of the experienced analyst which leads to the correct sequence and identification.

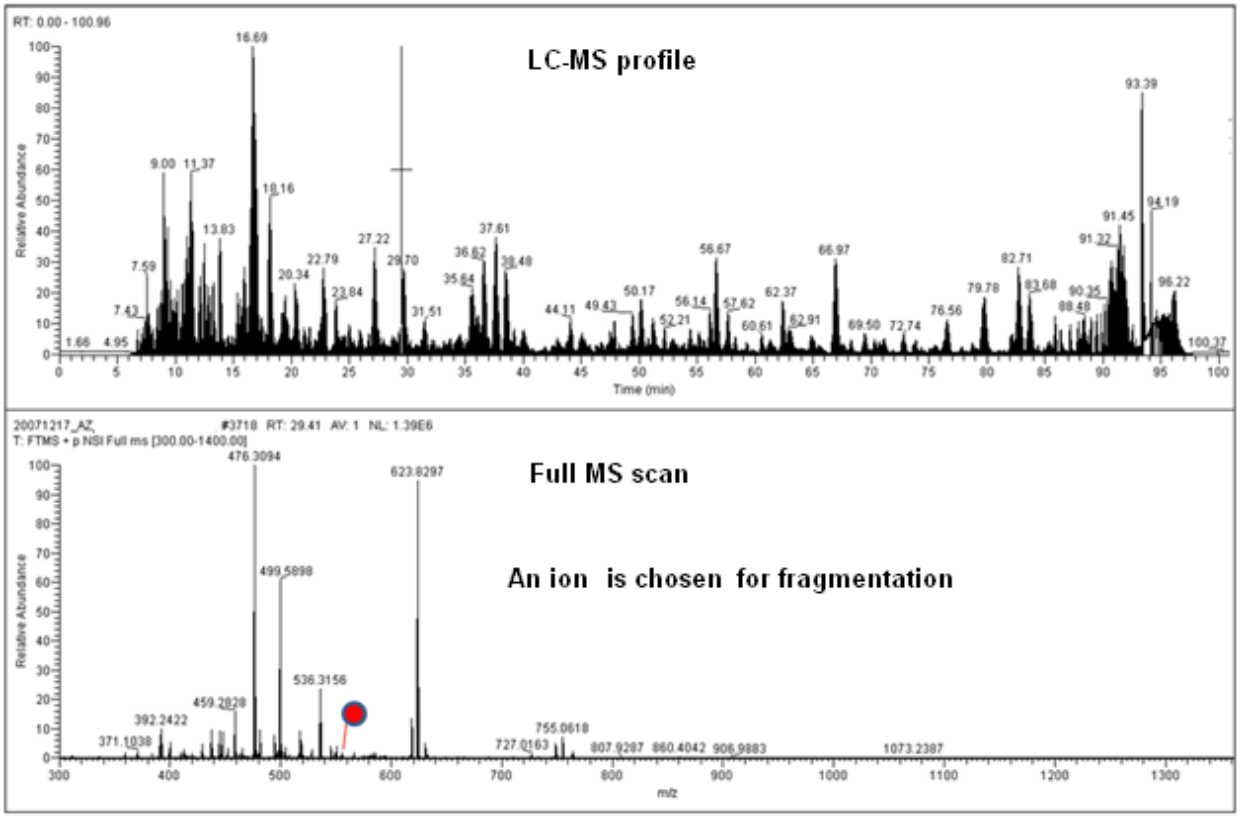
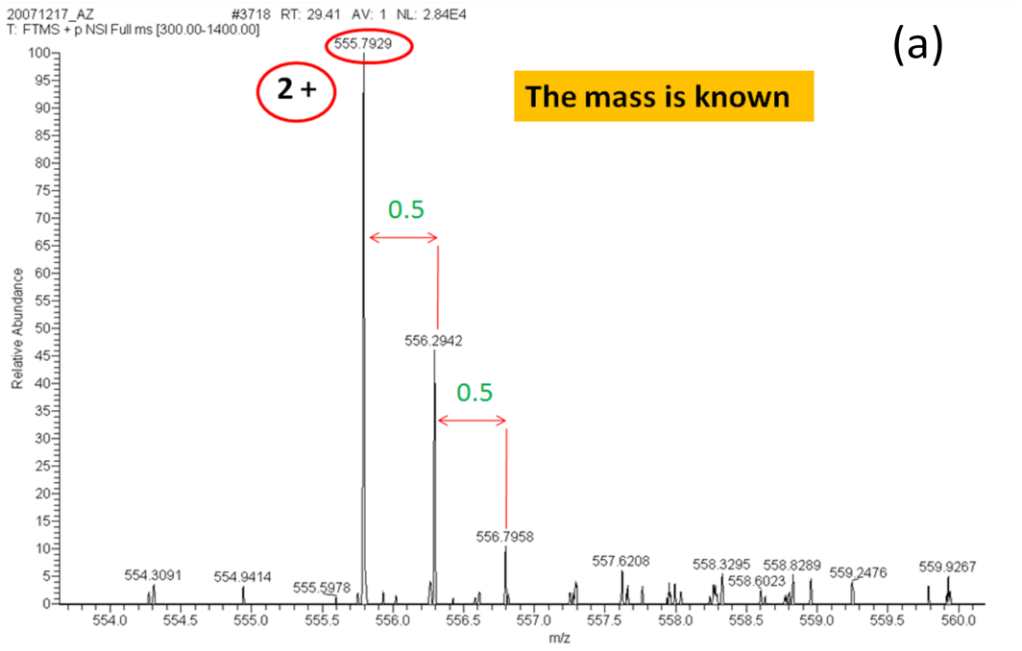


Figure 14. LC-MS profile of a complex peptide mixture. During the MS scan ions are chosen for isolation and the following MS/MS fragmentation.



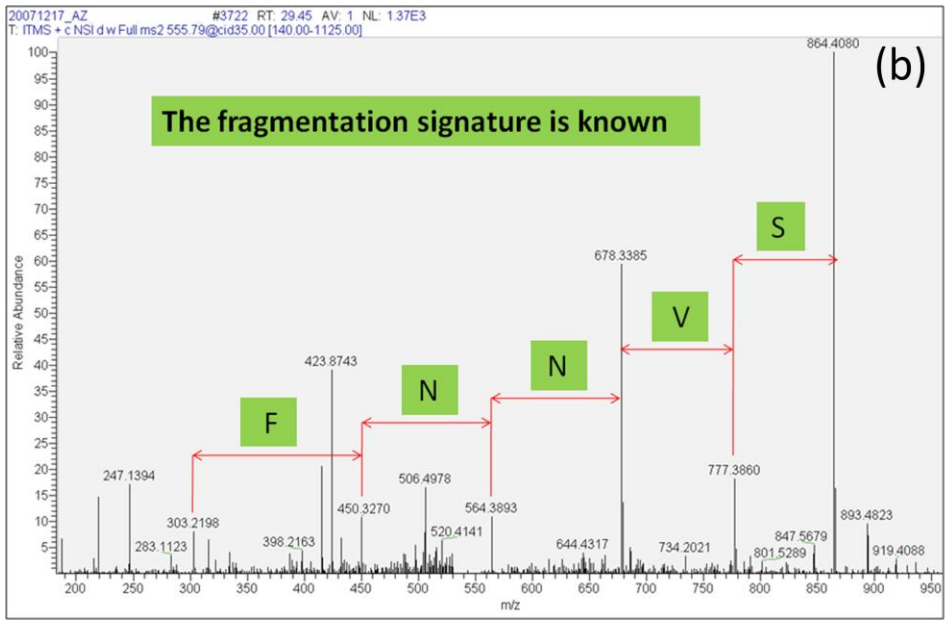


Figure 15. Information obtained during MS and MS/MS scans is used for peptide identification. (a) The peptide mass is calculated during the MS scan, (b) peptide internal composition is revealed during MS/MS scan.

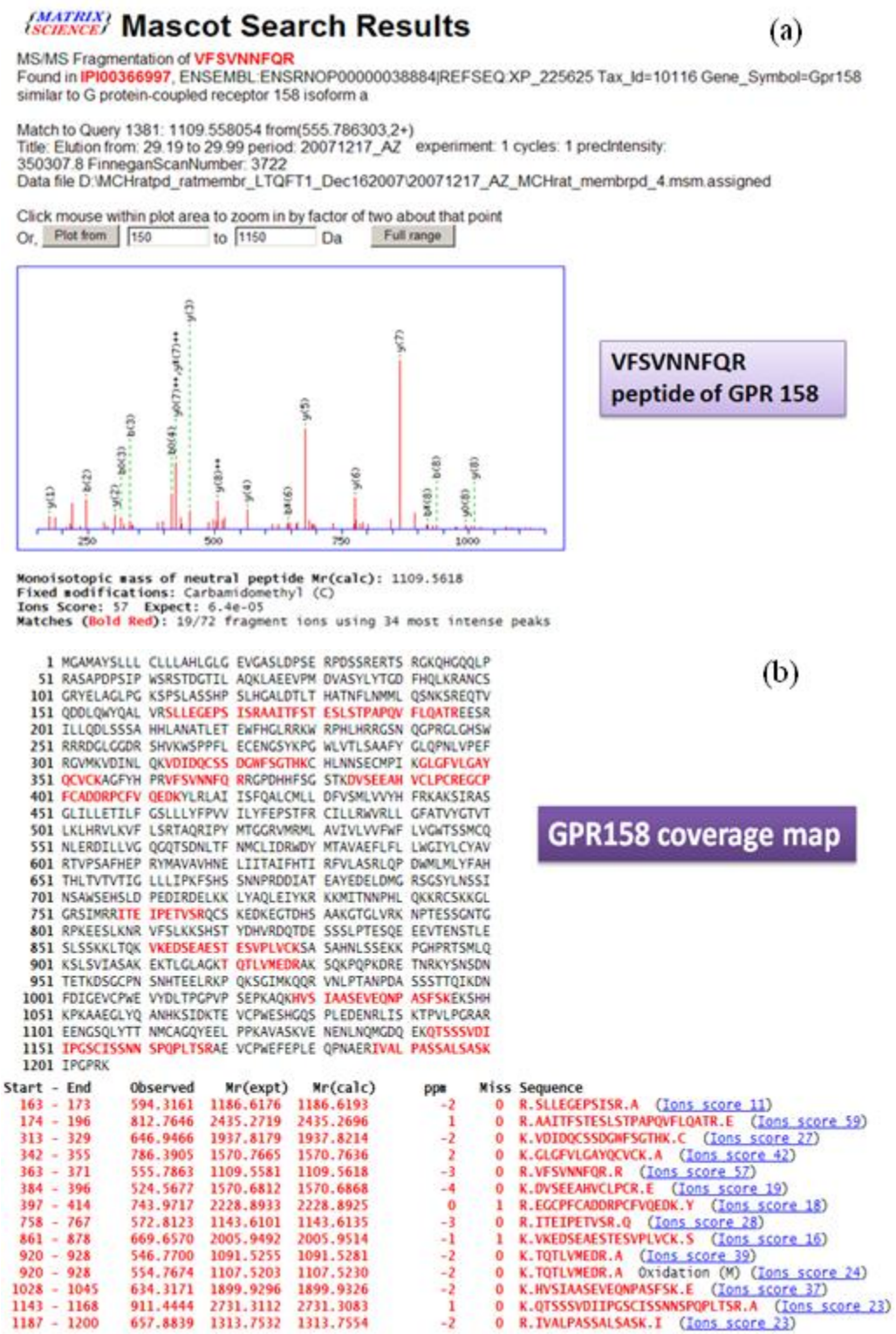


Figure 16. Identification output of a protein identification software engine. Single peptide identifications (a) are added up to create a protein peptide coverage map (b).

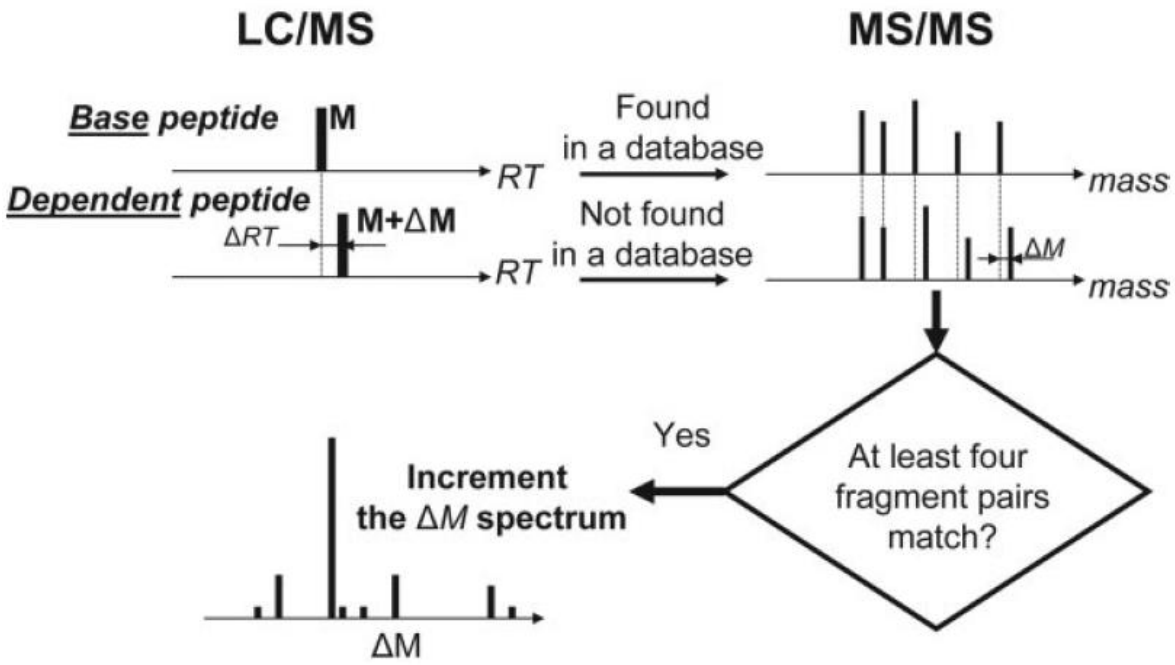


Figure 17. Modificomb algorithm (from ref. (*38*)).

The quadrupole-time-of-flight mass spectrometer (Q-TOF)

In my opinion, the Q-TOF mass spectrometer *QSTAR* introduced in the late 1990s by SCIEX was the first robust instrument generating high quality data obtainable in both MS and MS/MS modes which provided mass spectrometrists with the needed precision tools for investigating peptide structure and for *de novo* sequencing. Even though not as fast and accurate as the recently introduced *orbitrap*, the instrument was highly useful and popular in the proteomics community in the early 2000s. An excellent Q-TOF tutorial was written by Igor Chernushevich of SCIEX, one of the instrument's inventors, in 2001 (*9*). The Q-TOF tandem mass spectrometer of SCIEX can be described as a triple quadrupole with the detecting quadrupole section replaced by a TOF analyzer. An additional RF quadrupole Q0 is added to provide collisional damping, so the instrument consists of three quadrupoles, Q0, Q1 and Q2, followed by a reflecting TOF mass

analyzer with orthogonal injection of ions (Figure 18). The instrument combines filtering and MS/MS capabilities of the quadrupole analyzer with the accurate 20-50 parts per million (ppm) mass accuracy for detection in the TOF analyzer. For MS scans the quadrupoles operate in the radio frequency (RF) only mode, guiding ions to the TOF analyzer for detection. For MS/MS, Q1 is operated in the mass filter mode to transmit only the precursor ions of interest. The ions are then typically accelerated to between 25 and 50 eV before entering the collision cell Q2, where they undergo CID after collisions with neutral gas molecules. The resulting fragment ions are detected in the TOF part. In TOFMS, ions are resolved according to the formula:

$$R_{\mathrm{FWHM}} = \frac{m}{\Delta m} = \frac{t}{2\Delta t} \approx \frac{L_{\mathrm{eff}}}{2\Delta z}$$

where m and t are mass and flight time of the ion, Δm and Δt are the peak widths at the 50% level of the mass and time, respectively, Δz is the width of an ion packet in the vicinity of the detector and L_{eff} is the effective length of the TOF analyzer. When R is determined from t, there is an adverse factor of 2 reducing the resolution, which originates from the square root dependence of the flight time on mass:

$$t = \frac{L_{\mathrm{eff}}}{\sqrt{2eU_{\mathrm{acc}}}}\sqrt{m/z}$$

where U_{acc} is the full accelerating voltage in the TOF.

A typical Q-TOF sequencing cycle would include 1 MS scan (1 sec) and 3 MS/MS (2-3 sec each) which amounts to 7-11 sec acquisition time. Even though this is significantly slower than the cycle time of ion trap mass spectrometers, the Q-TOF provides high quality data in both MS and MS/MS modes which improves the quality of protein identification, and undoubtedly stimulated future spectacular developments in high accuracy biological mass spectrometry.

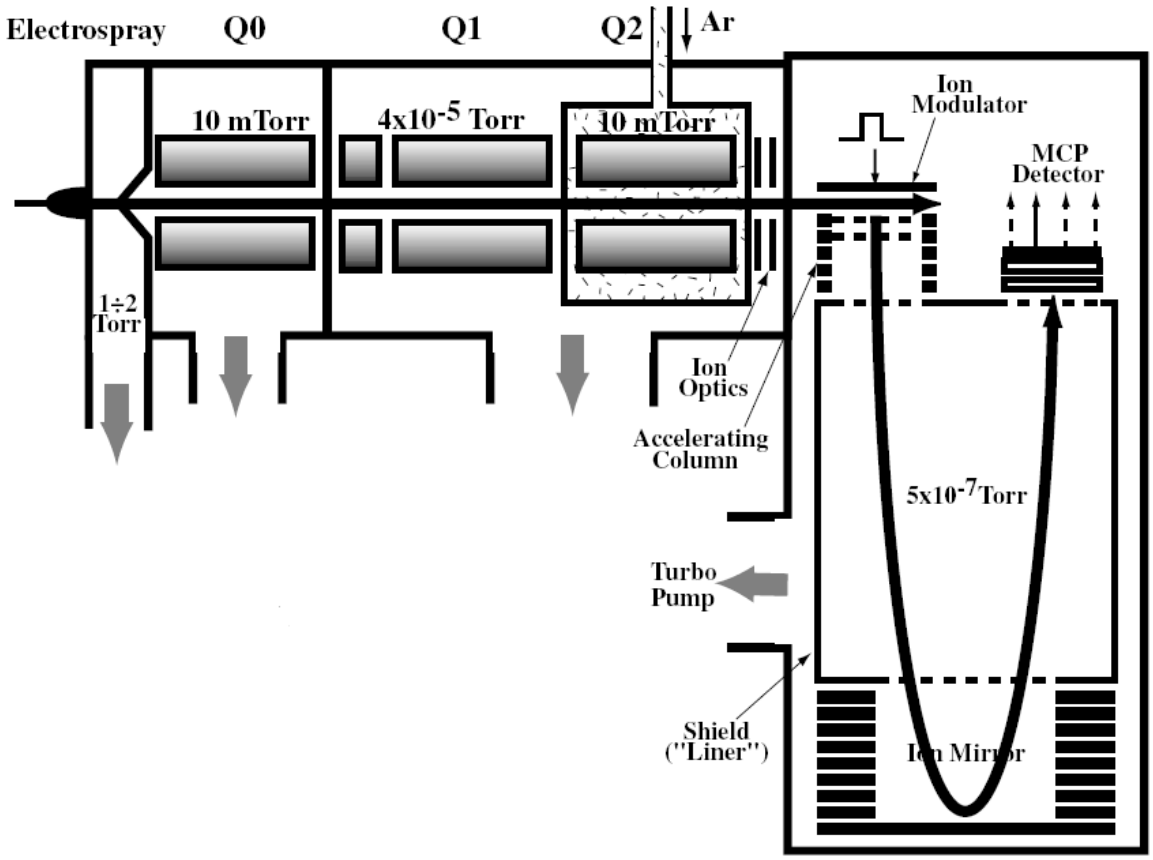


Figure 18. SCIEX QSTAR Q-TOF mass spectrometer (from ref. (*9*)).

The Orbitrap

The orbitrap analyzer is an example of how an old concept, if resurrected at the right time by an exceptional inventor, can lead to the creation of a truly revolutionary device. As mentioned before, orbital trapping was originally introduced by Kingdon in 1923 (*10*). The Kingdon trap contains a wire extended along the axis of a cylinder. When a voltage is applied between the wire and the cylinder, the created field attracts externally introduced ions to the wire. Only the ions that have enough tangential velocity do not collide with the wire and survive - electrostatic attraction is compensated by centrifugal force arising from the initial tangential velocity. The ions start orbiting around the wire in a manner similar to planets orbiting around their sun. The shape of the outer cylinder creates a field curvature which restrains axial motion of the ions. The so-called "ideal" orbital trap was introduced about 50 years ago (*42-44*) – its outer cylinder electrode is elaborately shaped and ions are trapped in the field with the potential distribution:

$$U(r,z) = \frac{k}{2} \cdot \left(z^2 - r^2/2 + R_m^2 \cdot \ln(r/R_m) \right)$$

where *r* and *z* are cylindrical coordinates , *k* is the field curvature, and *Rm* is the characteristic radius (Figure 19).

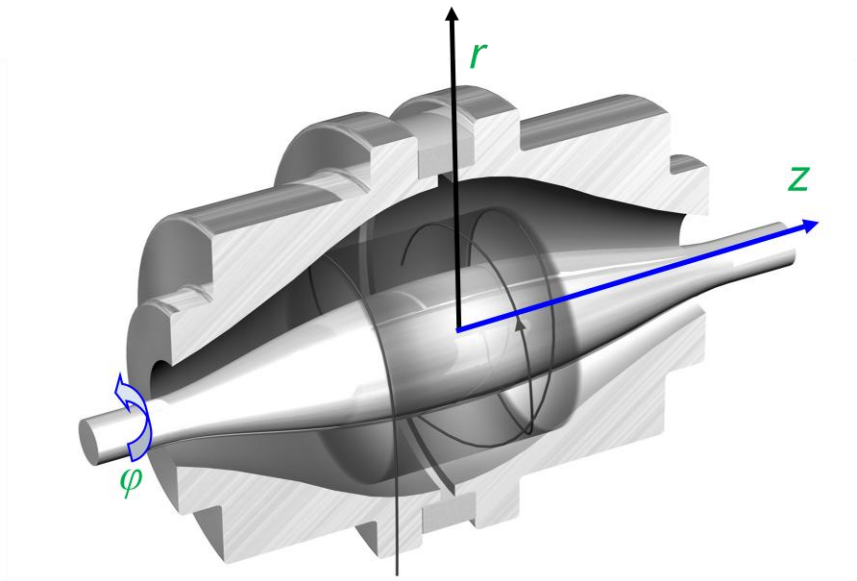


Figure 19. An ideal Kingdon trap (from ref. (*45*)). Ions are moving in spirals around the central electrode.

Before the patent by Makarov (*46*) it was proposed to use the "ideal" Kingdon trap as a mass spectrometer with image current detection (*47*). However, the measurement of m/z ratios was based on the frequencies of ion rotations. This approach led to poor mass resolution because ion velocity and initial radius significantly influence the rotational frequency. Makarov's idea was to derive the m/z values from the frequency of harmonic ion oscillations *along the* field's axis - this frequency is completely independent of energy and spatial spread of ions. The axial frequency is detected by processing the image current with Fourier Transform (FT) algorithms.

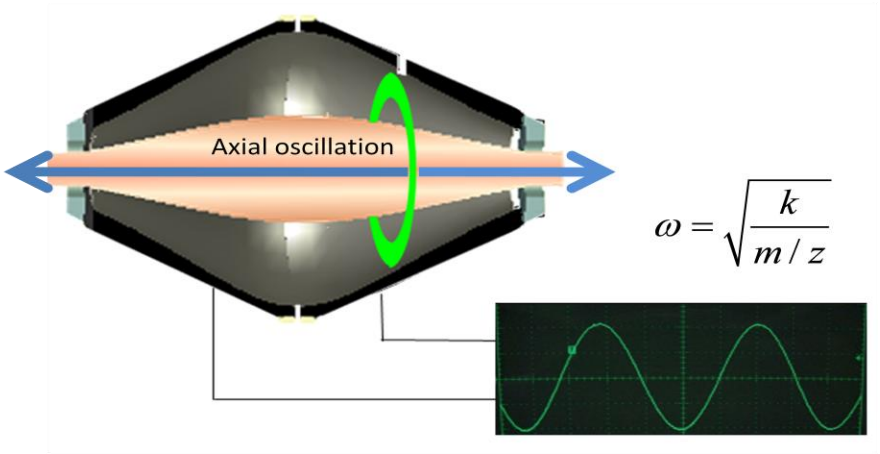


Figure 20. Axial oscillations of ions in the orbitrap do not depend on initial energy, angle and position of ions . ω is the oscillation frequency and k is an instrumental constant. Adapted from ref. (*45*).

The orbitrap measures mass with a very high accuracy of about 2-3 ppm. In addition to that, introduction of the so-called "lock-mass" option utilizing common ambient air contaminants (siloxane derivatives), which become ionized during electrospray, as internal calibration standards routinely increases mass measurement accuracy to below 1 ppm (*48*). In order to achieve its maximum performance, the orbitrap has to operate at the very low pressure of 10^{-10} mbar. For this reason it is not practical to fragment the ions inside the orbitrap by the frequently used CID process which relies on much higher pressures of 10^{-3}-10^{-5} mbar. The LTQ-Orbitrap hybrid instrument links the linear quadrupole ion trap (LTQ) (*49*) in tandem with the orbitrap. The linear ion trap is very sensitive and fast mass spectrometer used for guidance, storage, fragmentation and detection of ions even though it has relatively low resolution and accuracy. For peptide identification, mass accuracy of just 1 ppm for the precursor ion constrains peptide candidates to just a few sequences in the database, resulting in reduction of the false positive identifications, and, hence, greatly facilitates peptide identification (*50*). In the typical

proteomics set-up the precursor peptide masses are measured in the orbitrap (MS scan) with very high accuracy and the following fragmentation profiles are acquired in the LTQ with low resolution (MS/MS scans). The process is fast and, potentially, allows parallel operation which could lead to a rate close to 1 sec per cycle (1 cycle contains 1 MS scan and 5 MS/MS scans) (*51*). The obtained information is typically sufficient to identify a peptide product of a known protein in the database. Nevertheless, the quality of the MS/MS data obtained in the ion trap is not normally satisfactory if the need for manual *de-novo* interpretation of data arises. In this case, we can also perform fragmentation of the ions in the ion trap and forward the products into the orbitrap for high accuracy mass detection. Certainly the resulting data could be used for *de novo* analysis, however the design of the ion traps imposes a low mass cut-off restriction on the detectable product masses which usually corresponds to 1/3 of the precursor ion mass**.** Thus, such important MS/MS pieces of evidence as immonium, a_2 and b_2 ions (*23*) are not present, which could impede elucidation of peptide structure.

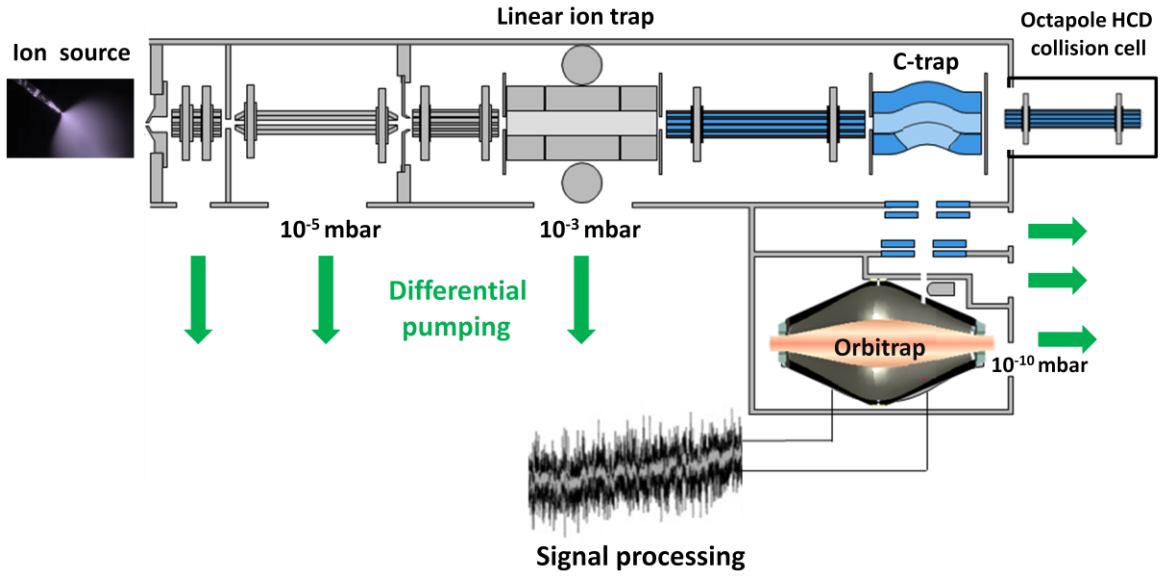


Figure 21. LTQ-Orbitrap XL mass spectrometer (modified from ref. (*52, 53*).

Higher-energy C-trap dissociation (HCD)

Recently, Olsen et al. showed that ions can be efficiently fragmented by high-accuracy and full-mass-range tandem mass spectrometry (MS/MS) with a method termed higher-energy C-trap dissociation (HCD) (*53*). The C-trap, normally used to store ions on their way from the ion trap to the orbitrap, can also be used as a collision chamber to enable low energy CID fragmentation which yields y and b ion series typical for quadrupole-type CID. In this case, the fragmentation spectra are obtained over the full mass range and provide an analyst with useful information about the low m/z reporter ions. In the LTQ Orbitrap XL, the newest version of the LTQ-Orbitrap line, fragmentation is performed in a separate octapole collision cell at the far end of the C-trap. The octapole collision cell is aligned with the C-trap. The cell contains nitrogen at low pressure as a collision gas. The collision cell is supplied with an RF voltage of which the direct current offset can be varied thus giving the opportunity of changing the values of collision energy (CE).

The following example illustrates how adjusting the applied CE can be helpful for elucidation of post-translational modifications. O-linked sugars are very labile and can easily be destroyed during the mass spectrometric fragmentation event. Figure 22 presents a fragmentation spectrum of a naturally occurring glycosylated peptide product of pro-IGF2 from our cerebrospinal fluid (CSF) profiling study (*54*). This peptide, with sequence *DVSTPPTVLPDNFPRYPVGKF*, is modified by O-linked glycosylation at T-99 and carries the $HexNAcHex(NeuAc)_2$ sugar. When the HCD fragmentation spectrum of the pro-IGF2- derived peptide was acquired with the CE value of 50, we observe a singly charged ion at *m*/*z* 274.092 corresponding to the dehydrated oxonium ion of sialic acid. We can also clearly follow the *y*-ion series, and bearing in mind the presence of a potential sugar attachment, we identified the peptide. Nevertheless, the information about the exact location of the modification, as well as its structure, is lost at this CE value. We found that by lowering the CE from 50 to 35 we did observe fragment ions pertaining to sugar attachments (Figure 23). In this case, the sugars are partially preserved which enables us not only to identify the O-linkage site but also to ascertain the glycosylation structure. As the direct result of this identification, we modified the MASCOT search engine to take the characteristic mass offsets of the discovered glycosylation signature into account. In this way we discovered some

novel neuropeptides solely as their glycosylated forms which, otherwise, would never have been found (*54*).

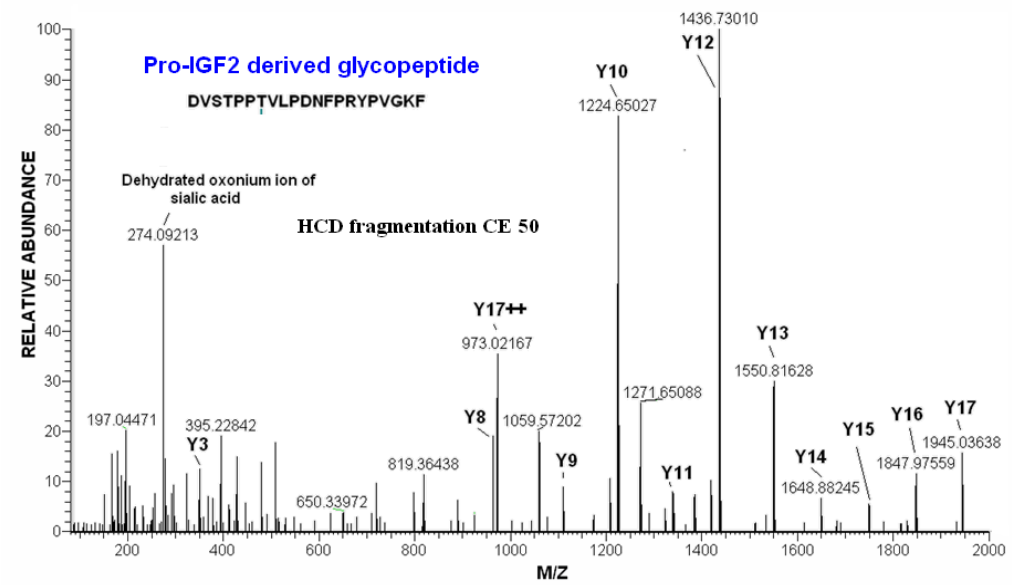


Figure 22. *DVSTPPTVLPDNFPRYPVGKF* glycopeptide, HCD fragmentation at CE of 50.

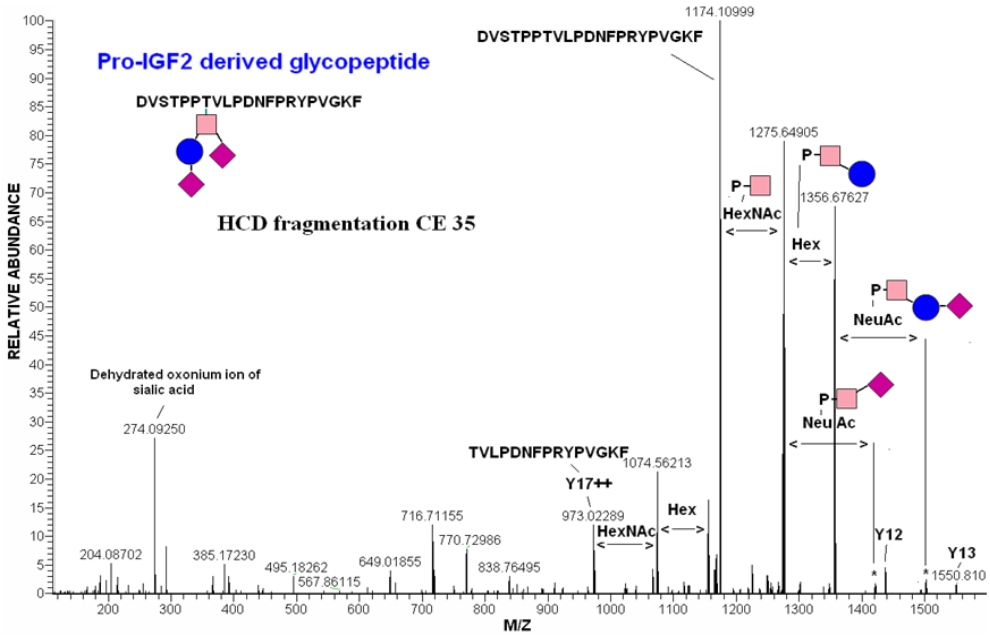


Figure 23. *DVSTPPTVLPDNFPRYPVGKF* glycopeptide, HCD fragmentation at CE of 35.

De-novo sequencing as a useful proteomics tool

I consider peptide de-novo sequencing to be an extremely useful proteomics tool. I was trained as a biological mass spectrometrist about 10 years ago when search engines for mass spectrometric peptide data were yet to be made reliable and the accuracy of MS instrumentation was yet to improve. At that time, most of the peptide identifications demanded verification by experts, and, as a consequence, the time provided an excellent opportunity to learn the intricacies of MS/MS and the quintessence of *de-novo* peptide identification. Nowadays, sophisticated search engines provide answers to many submitted enquiries. Nevertheless, *de-novo* sequencing is still an indispensible tool for structure elucidation of unknown proteins and peptides, as well as characterization of post-translational modifications. In the presented publications, I used *de-novo* sequencing on a high accuracy and resolution mass spectrometer for identification and characterization of novel CSF neuropeptides, their post-translational modifications (*54*) as well as discovery of novel extended forms of nuclear proteins (*55*).

Proteomics in genomics

The Human Genome Project was completed in 2003 and provided scientists with the means for prediction of the protein-coding genes. Until recently, a lot of effort has been spent on the development of the bioinformatics-based tools for gene prediction and annotation. For mammalian genome annotation the Ensembl (*56*) and NCBI (*57*) are the major processing gateways. Both cDNA-derived, evolutionary-derived and *de novo* approaches are used for the prediction of the protein-coding gene sequences. Even though eight years of bioinformatics progress have gone by since the Bork's publication in which he described the pitfalls of the gene prediction algorithms estimating the false positive rate of gene prediction to be at least 30% (*58*), his statement is still relevant nowadays. Obviously, no gene prediction pipeline is 100% fault-safe. For the human genome, for example, *de novo* gene prediction approaches are estimated to be only 50% correct (*59*). High frequency of alternative splicing in mammalian genomes is one of the major contributors to erroneous gene assignments. For the verification of the predicted genes, the expression-based techniques are used most commonly. The techniques involve RT-PCR and direct sequencing of predicted protein-coding genes. While these methods can validate

the expression of the predicted gene, they do not provide an answer as to whether the expressed gene is translated into a protein. Proteomics identifies protein products of the genes and reports on the "real" output of the translation machinery. Hence, in addition to confirming predicted genes, proteomics can also supply information leading to discovery of novel genes, splice and edited gene variants. Genomics and proteomics are truly complementary of each other and, no doubt, in the future we will see the fusion of the two fields. A number of the recent initial studies reported on using proteomics tools to probe the genome (*60-62*). One of the most obvious ways to fish out novel genes from LC-MS/MS data is to perform a search against an EST or gene database. However, the queried database must be translated in six frames which, giving the complexity of the human genome, results in a huge database size (for comparison – a "normal" human protein sequence database has the size of about 40 MB, the translated EST human database - 7 GB). Additionally, the size of the processed LC-MS/MS files obtained by contemporary high-throughput instrumentation is also massive and could easily go over a few gigabytes per proteome. This creates significant computational challenges for LC-MS/MS-based identification of novel protein-coding genes from genomic databases. The key need is currently for innovative computational approaches which will help to overcome the above-mentioned obstacles.

Neuropeptidomics

The neuropeptides are naturally occurring 3-100 amino-acid residues long polypeptides. It is now a common knowledge that neuropeptides, similarly to such classical neurotransmitters as amino acids, metabolites and biogenic monoamines, can target specific receptors and induce a wide spectrum of physiological responses. The first neuropeptide, substance P, was discovered by von Euler and Gaddum in 1931 (*63*). However, the amino acid composition of substance P was revealed only fifty years later when Chang and Leeman described cloning and sequencing of substance P (*64*). They showed that substance P is 11 amino-acid long peptide and C-terminally amidated. Neuropeptides are smaller than average proteins and present both in CNS and peripheral organs, their interaction with the target receptors is much stronger than that of small neurotransmitters (*65*). During a typical neuropeptide processing event, the signal peptide is removed from a protein precursor by a signal peptidase (*66*). The pro-peptide is cleaved by

specific convertases at dibasic sites generating precursors which then are acted upon by carboxypeptidases (*67*) giving rise to mature forms of neuropeptides. In addition to this, the peptides could be C-terminally amidated by peptidylglycine monooxigenase which requires a C-terminal glycine in the peptide sequence (*68*). MS-based neuropeptidomics is a nascent and promising area of neuropeptide research (*69*). Even though recent developments in capillary separation science and improvements in sensitivity and accuracy of mass spectrometers have provided scientists with necessary tools for discovery of novel neuropeptides, the identification of neuropeptides by LC-MS/MS is still challenging. In conventional proteome mapping experiments, proteins are digested with trypsin, and each protein is usually identified by at least two fully tryptic and unmodified peptides. Tryptic peptides which carry C-terminal basic amino acids are protonated under acidic conditions – the event facilitates their mass spectrometric fragmentation and consequent interpretation of the data. The identification of neuropeptides is much more challenging. They do not terminate in predictable amino acids, making the number of candidates to be considered in the database search much higher, and the nonstandard charge distribution within the peptide sequence can produce fragmentation spectra with a smaller number of characteristic ions. Importantly, neuropeptides can be post-translationally modified which additionally complicates their analysis. In many instances, researchers still have to rely on manual interpretation of the neuropeptide MS/MS fragmentation patterns. It is obvious that there is a huge demand for improved efficient computational platforms for the LC-MS/MS neuropeptide identification.

Presented work

Proteomics as a tool for discovery of novel genetic editing mechanisms

There is a widespread belief that thorough analysis of unassigned shotgun proteomics data could lead to discovery of novel genes, novel post-translational modifications or novel gene expression events (*70*). The data quality obtainable with modern high accuracy mass spectrometry greatly facilitates the attempts to 'fish out' something novel from the depths of the proteomic. However, until now three were almost no examples of useful biological data emerging from such attempts. Here, we used such information to discover a previously unknown mechanism creating altered protein forms (*71*).

Linker histones H1 and high-mobility group (HMG) proteins are abundant nuclear proteins that regulate gene expression through modulation of chromatin structure. For a number of years, we have analyzed these proteins and have reported on their modifications in cell culture and in tissues (*72, 73*). A while ago we noticed that many fragmentation spectra of excellent quality could nevertheless not be mapped to any protein in the database. We then performed 'de novo' sequencing on these peptides (Figure 24) and found, to our surprise, that they mapped just upstream of their respective genes.

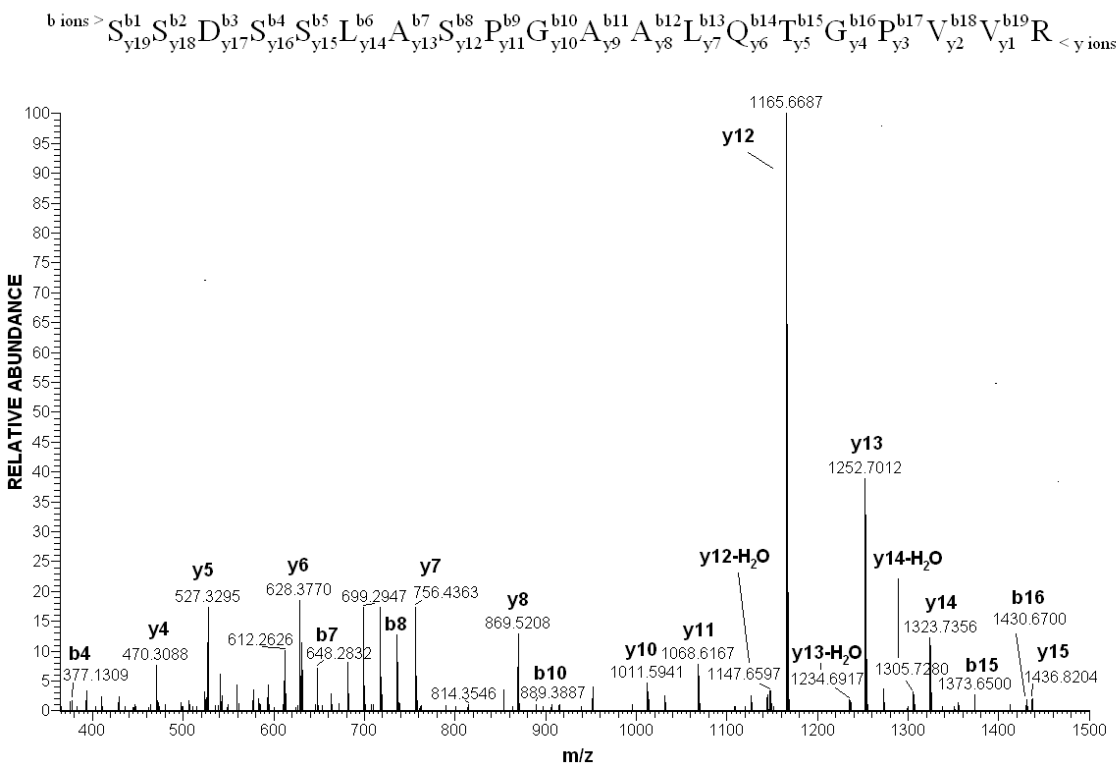


Figure 24. High accuracy MS/MS spectrum of the SSDSSLASPGAALQTGPVVR peptide pertaining to the 5'-UTR of *h1.0*. De-novo sequencing of the peptide provided a stimulus for further research and lead to discovery of the previously not reported insertion editing mechanism in humans.

This was not due to an alternative, upstream start codon because the 5'UTRs do not contain any such stop codon. We then analyzed expressed sequence tag (EST) databases and found that some EST sequences contained an additional, inserted uridine, creating a novel AUG start site.

Predicted ET part of H1.0

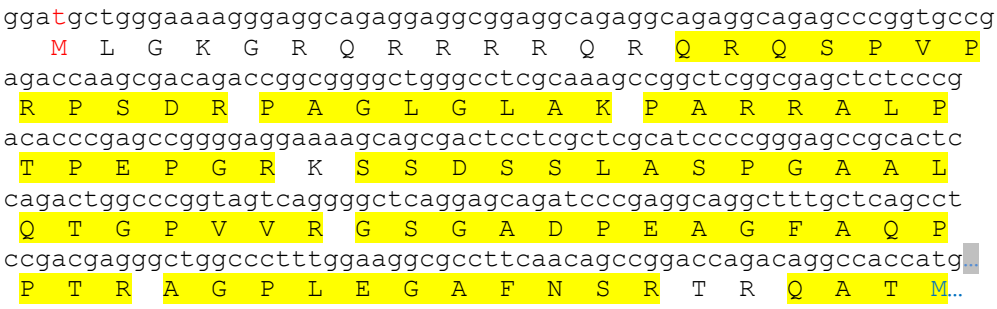


Figure 25. Sequence coverage of the extended (ET) part of H1.0.

While RNA insertion has not been described in metazoans, it is not completely unprecedented in biology and may resemble the RNA editing mechanism described for trypanosomatid protozoans in which the uridine insertion/deletion-based RNA editing of the mitochondrial mRNA creates new initiation/termination codons (*74*). During the last years we have raised antibodies against these proteins and characterized their localization and behavior. Interestingly, we find that they are regulated differently from their 'parent genes' - the extended (ET) form of histone H1, for example, found in splicing speckles, is upregulated upon apoptotic treatment with vinblastine and TNF-alpha, whereas H1 itself is not affected by the treatment. The N-terminally extended proteins occur in normal human cells and their synthesis is not related to alterations at the DNA level. For the described examples, the amounts of ET-proteins appear to be about two orders of magnitude lower than those of the standard products of the corresponding genes. However, since H1.0 and HMGN1 occur in about 10^6-10^7 copies per cell the ET forms are still relatively abundant compared to transcription factors and other chromatin proteins with specific regulatory functions. Presumably, the ET-proteins were not identified until now because they are not predictable from the corresponding genes and they occur 'in the shadow' of their normal forms. Thus, in the past, the ET-proteins were probably observed in many experiments but simply ignored as artifacts - as evident from Figure 26 commercial antibody against HMGN1 recognizes both "normal" and the extended versions of this protein.

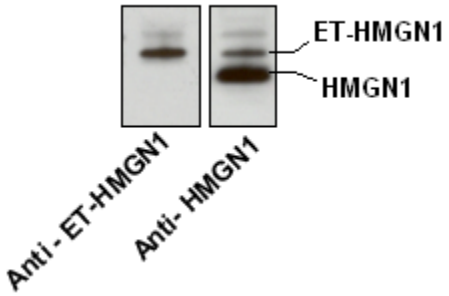


Figure 26. Western blots of perchloric acid extracts of MCF7 cells probed with antibodies against ET-HMGN1 and HMGN1, respectively.

Current Biology *18*, 1760–1765, November 25, 2008 ©2008 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2008.09.059

Report

Evidence for Insertional RNA Editing in Humans

Alexandre Zougman,[1,3] Piotr Ziółkowski,[2] Matthias Mann,[1,*] and Jacek R. Wiśniewski[1,*]
[1]Department of Proteomics and Signal Transduction
Max Planck Institute for Biochemistry
Am Klopferspitz 18
D-82152 Martinsried
Germany
[2]Department of Pathology
Wroclaw Medical University
ul. Marcinkowskiego 1
PL-50-368 Wroclaw
Poland
[3]Center for Integrated Protein Science
D-81377 Munich
Germany

Summary

Large-scale analysis directly at the protein level holds the promise of uncovering features not apparent or present at the gene level [1–3]. Although mass spectrometry (MS)-based proteomics can now identify and quantify thousands of cellular proteins in large-scale proteomics experiments, much of the peptide information contained in these experiments remains unassigned [4]. Here, we use such information to discover a previously unreported mechanism creating altered protein forms. Linker histones H1 and high-mobility group (HMG) proteins are abundant nuclear proteins that regulate gene expression through modulation of chromatin structure [5–8]. In the high-resolution MS analysis of histone H1 and HMG protein fractions isolated from human cells, we discovered peptides that mapped upstream of the known translation start sites of these genes. No alternative upstream start site exists in the genome, but analysis of Expressed Sequence Tag (EST) databases revealed that these N-terminally extended (ET) proteins are due to in-frame translation of the 5′ untranslated region (5′UTR) sequences of the transcripts. The new translation start sites are created by a single uridine insertion between AG, reflecting a previously unreported RNA-editing mechanism. To our knowledge, this is the first report of RNA-insertion editing in humans and may be an example of the type of discoveries possible with modern proteomics methods.

Results and Discussion

Identification of N-Terminally Extended (ET) Proteins

In-depth proteomic analysis of nuclear extracts from various human cell lines revealed a multitude of posttranslational modifications in linker histone H1 and high-mobility group (HMG) proteins [9]. Apart from this, we found a number of fragmentation spectra of excellent quality that could not be matched to any protein sequence or open reading frame (ORF) in the human International Protein Index (IPI) database. We therefore extracted partial de novo sequences from these fragmentation spectra. When searching the NCBI dbEST database [10] with these sequences, we found that they matched directly upstream of the start codon and within the 5′UTR of *h1.0.* Similarly, we also found peptides pertaining to the 5′UTR of the *hmgn1* gene. This was surprising because in both cases, the 5′UTRs do not contain any alternative start codons.

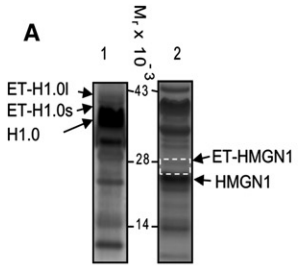
To characterize the N-terminally extended (ET) sequences of H1.0 and HMGN1, we partially purified these proteins from MCF7 cells by reversed-phase chromatography, resolved fractions by SDS PAGE (Figure 1A), digested them with trypsin, and analyzed peptide mixtures by online liquid-chromatography mass spectrometry (LCMS). We identified a total of eight unique peptides (including one phosphopeptide) of ET-H1.0 and two unique peptides of ET-HMGN1 using high-resolution, high-mass-accuracy MS. Several of these peptides were further verified by high-resolution fragmentation analysis with low ppm accuracy (see Appendices S1 and S2, available online). Fragmentation spectra of the identified peptides containing the "normal" initiation methionine, as well as the peptides most proximal to the N termini of ET-H1.0 and ET-HMGN1, are shown in Figure 1. We identified two forms of ET-H1.0, visible as two bands in SDS PAGE, and a single form of ET-HMGN1 (Figure 1F). On the basis of the mapped peptides, the long and short ET extensions of H1.0 comprise at least 86 and 63 amino acid residues, respectively, whereas the ET extension of HMGN1 is at least 23 amino acid residues in length. Due to their amino acid composition, the high number of the basic residues, and the low number of the hydrophobic residues, histones have reduced mobility in SDS PAGE. With respect to the molecular weight (MW) markers, H1 histones appear at a MW of 30,000–35,000 Da. However, their molecular masses are in the range of 20,000–22,000 Da. This is the reason that the ET-H1.0 forms appear to have MWs close to 40,000 Da.

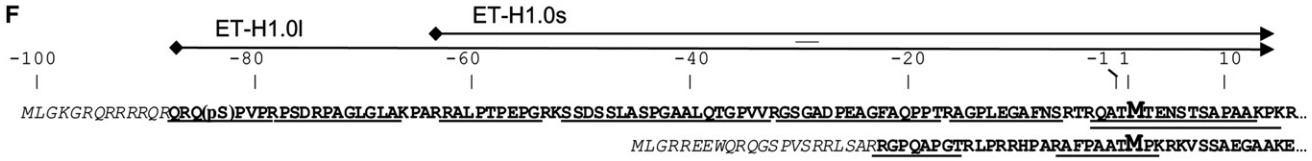
New Translation Start Sites Are Created by a Single Uridine Insertion

We then searched the NCBI dbEST database for EST fragments containing the 5′UTR of *h1.0* and found 11 sequences, derived from different tissue sources, with a uridine (U) insertion (the NCBI dbEST database release of August 06, 2008 contains 301 ESTs carrying the *h1.0* 5′UTR, 11 of them with U insertion). In each case, this insertion happened at the same position—297 bp upstream of the known start codon—between the A and G bases at the AGCT location (Figure 2B, Appendices S3, S4, and S5). For *hmgn1*, we likewise found an EST sequence in which a new start codon is potentially created by U insertion, in this case 135 nt upstream of the known start codon (Appendices S3 and S5). With these start sites, the identified peptides cover 85% and 50% of the predicted sequences of ET-H1.0 and ET-HMGN1, respectively (Figure 2A).

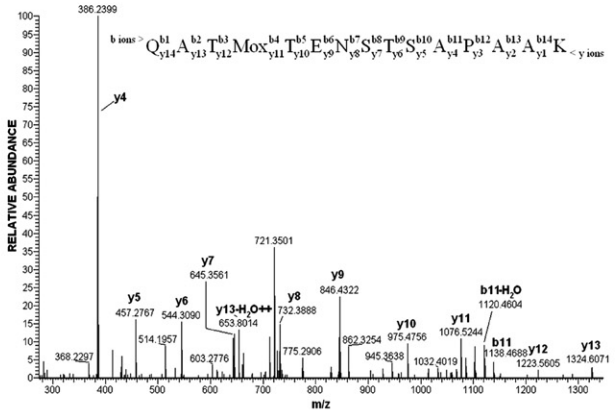
h1.0 occurs as a single gene in the genome database. To exclude the existence of a genomic form of ET-*h1.0*, we performed PCR analysis of genomic DNA for the sequence suggested by the EST data. The "normal" *h1.0* gene has an *AluI* restriction site (5′-AG/CT-3′) in its 5′UTR sequence at the position where the U insertion was found in some EST sequences

*Correspondence: mmann@biochem.mpg.de (M.M.), jwisniew@biochem.mpg.de (J.R.W.)

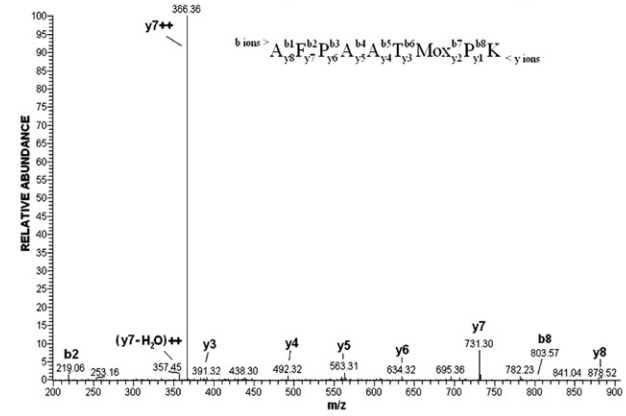




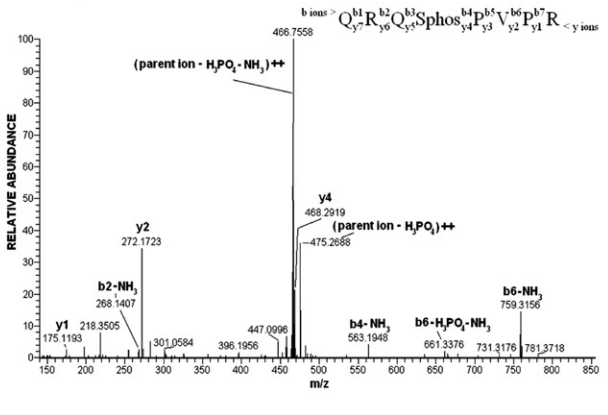
B Histone ET-H1.0 (residues -3−12)



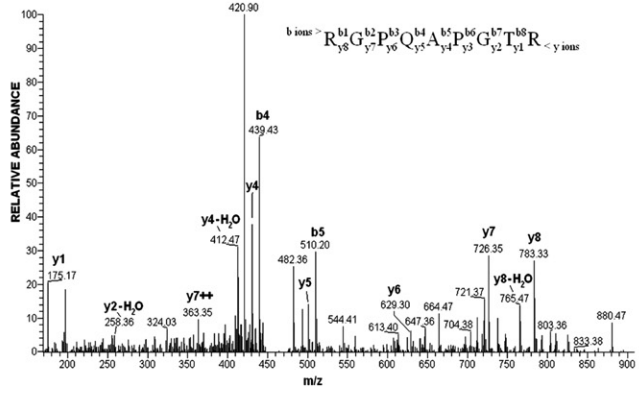
C High Mobility Group Protein ET-HMGN1 (residues -6−3)



D Histone ET-H1.0 (residues -86−-79)



E High Mobility Group Protein ET-HMGN1 (residues -23−18)



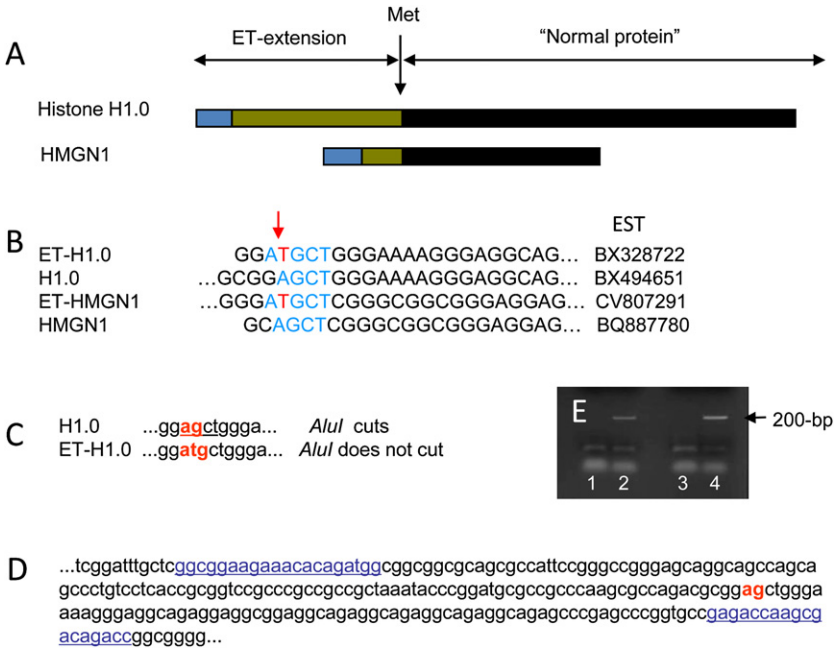


Figure 2. Primary Structure and Putative Origin of ET Proteins

(A) Schematic view of the proteins with extension of their termini (ET) originating from translation of the 5′UTRs. Black, canonical translation products; green, sequenced by high-resolution mass spectrometry; blue, predicted primary structure of the ET-H1.0 and ET-HMGN1 proteins.

(B) AGCT site with the T insertion (arrow) in EST sequences coding for ET proteins.

(C) The *AluI* site of the "normal" *h1.0* sequence is impaired by the creation of the new start codon.

(D) The 200 bp sequence containing the base insertion site (red bold) was amplified with genomic DNA isolated from MCF7 cells. The sequences of the used primers are underlined.

(E) The 200 bp PCR product was obtained only from *AluI*-undigested templates. Lane1, the DNA template was digested with *AluI* prior to the PCR reaction; Lane 2, the PCR reaction with intact template; Lanes 3 and 4, the 200-bp product from lane 2 was extracted from the gel and incubated in the presence or absence of *AluI*, respectively.

(Figure 2C). The insertion creates a new start codon and impairs the *AluI*-restriction site. Thus, amplification of the 200 bp sequence (Figure 2D) from *AluI*-treated DNA would allow identification of the modified fragment, and the PCR product obtained from the undigested template would be susceptible to *AluI* digestion. The PCR reactions revealed that neither was the 200 bp fragment amplified from the *AluI*-digested template nor was the 200 bp PCR product from the undigested template cleavable by the endonuclease (Figure 2E). These results disprove the presence of *h1.0* gene sequence coding for N-terminally extended H1.0, confirming that the new start codon is created during or after transcription of *h1.0* by an RNA editing process. It is possible that the *hmgn1* mRNA is processed by the same mechanism as that of *h1.0*. While RNA insertion has not been described in metazoans, it is not completely unprecedented in biology and may resemble the RNA editing mechanism described for trypanosomatid protozoans in which the uridine insertion/deletion-based RNA editing of the mitochondrial mRNA creates new initiation/termination codons [11]. The resulting N-terminal protein sequences can carry unique properties such as, for example, DNA binding [12]. The human 5′UTRs of H1.0 and HMGN1 have no obvious similarities with uridine insertion regions in the trypanosomatid mitochondrial genes. These regions also are not conserved between species.

ET Proteins Occur Abundantly in Human Cells

To characterize ET-H1.0 and ET-HMGN1 in human cells and tissues, we raised antibodies in rabbits against peptides derived from the N-terminal extensions and used the affinity-purified antibodies for western blot and immunofluorescence staining analyses (Figure 3). The western blot analyses revealed that both forms of ET-H1.0 were present in human breast- and lung-cancer cells but not in HeLa cells (Figure 3A). ET-H1.0 was also detected in extracts from three human-cancer tissues and three normal tissues (Figure 3A). Although ET-H1.0 was always accompanied by H1.0, its abundance did not correlate with the observed amounts of H1.0. For example, normal breast tissue and cancerous breast tissue from the same patient contained similar amounts of H1.0, whereas the abundance of the ET-H1.0 fluctuated (Figure 3A, lanes 4–9).

Immunofluorescence staining of MCF7 cells revealed a bright-speckled staining pattern of interphase nuclei (Figures 3B and 3C), and its staining pattern was distinct from that of H1.0 (Figures 3G and 3H). During mitosis, ET-H1.0 appears to be in the vicinity of condensed chromosomes (arrows in Figures 3B and 3C). Control staining of HeLa cells, in which we were not able to detect ET-H1.0 (see above), resulted in the absence of staining (Figures 3D and 3E). The preferential nuclear location of ET-H1.0 was confirmed by fractionation of MCF7 cells into nuclear and low-speed cytosolic fractions followed by western blotting (Figure 3G). Staining via a monoclonal H1.0 antibody was distinct from that via its ET form (Figures 3G and 3H).

We found that ET-H1.0 colocalizes with splicing speckles (Figures 3I–3L). The splicing speckles, subnuclear structures identified by immunofluorescence microscopy, are thought to mirror the interchromatin granule clusters detected by electron microscopy. The most widespread belief is that the

Figure 1. Identification of Linker Histone and HMG Proteins with Extra Terminal (ET) Extensions Created upon Translation of the 5′-Untranslated Regions

(A) Localization of ET-H1.0 and ET-HMGN1 in PAGE-separated nuclear protein fractions of MCF7 cells.

(B and C) Fragmentation spectra of tryptic peptides carrying the regions from both the ET extension and the "normal" protein parts of ET-H1.0 and ET-HMGN1, respectively.

(D and E) Fragmentation spectra of tryptic peptides carrying the most N-terminally identified sequences of the ET extension of ET-H1.0 and ET-HMGN1, respectively. See [25] for explanation of peptide fragmentation.

(F) Putative primary structure of the ET proteins. Bold text indicates experimentally observed ET extension; underlined text indicates sequences of indentified peptides. pS denotes phosphoserine. Negative numbering is used for amino acid residues in the ET extensions. Arrows indicate the primary structures of the long (l) and short (s) ET-H1.0 forms.

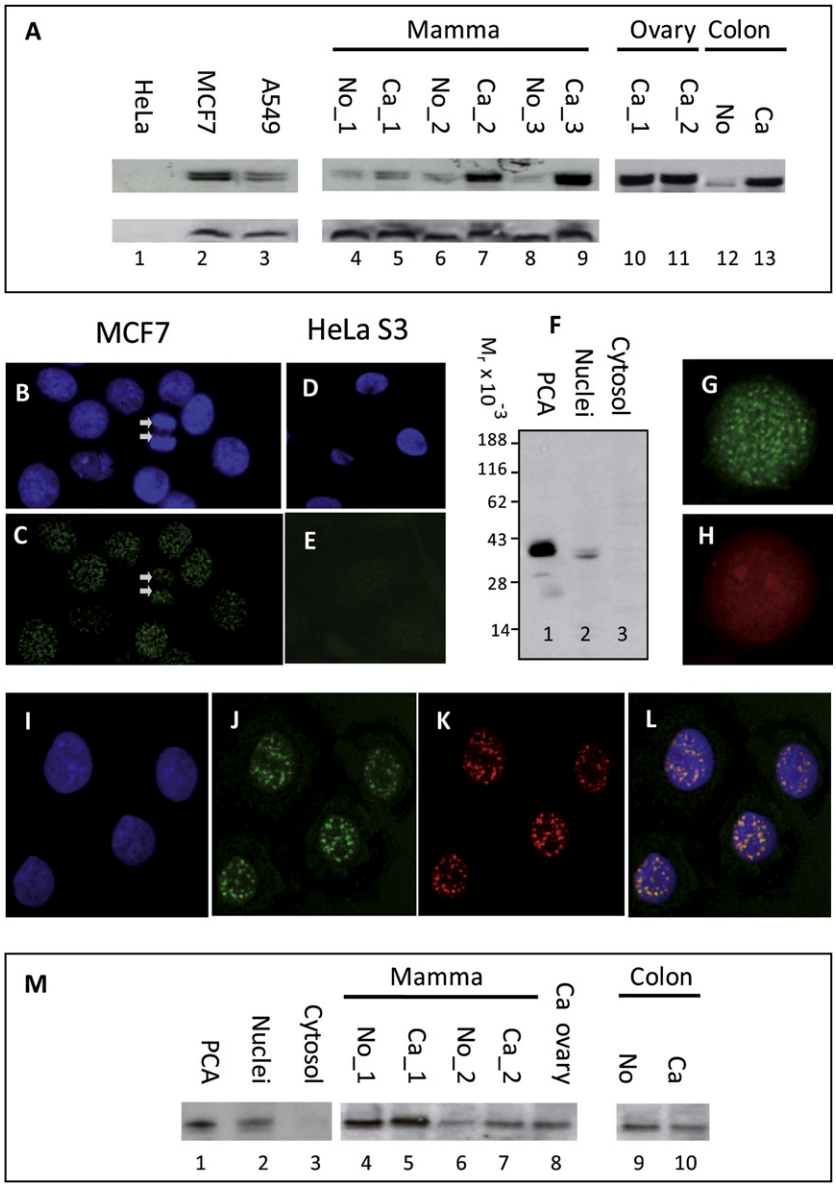


Figure 3. Occurrence of ET-H1.0 and ET-HMGN1 in Human Cells and Tissues

(A) Occurrence of ET-H1.0. Lanes 1–3, western blot analysis of perchloric acid extracts of cervical-, breast-, and lung-cancer cells, respectively. Lanes 4–9, extracts of three pairs of tissue, each matching the same patient (1, 2, or 3), of normal breast (No) and ductal invasive carcinoma G2 (Ca). Lanes 10 and 11, ovarian tumor. Lanes 11 and 13, normal colon and colon adenocarcinoma G2. The blots were probed with anti-ET-H1.0 and H1.0 antibodies.

(B–E) Nuclear localization of ET-H1.0 in MCF7 cells (B–D); absence of fluorescence in HeLa cells, which express little or no ET-H1.0 (E); fluorescent staining of nuclei with DAPI (B and D); and immunofluorescent staining with anti-ET-H1.0 antibodies (C and E).

(F) Western blot analysis of purified nuclei (lane 2) and cytosolic fraction (lane 3) of MCF7 cells. Lane 1, perchloric acid extract reference.

(G and H) Double immunofluorescent staining of the MCF7 nucleus with ET-H1.0 (green) and H1.0 (red) antibodies, respectively.

(I–L) Double immunofluorescent staining of the MCF7 nucleus with anti-ET-H1.0 (J) and SC-35 (K) antibodies, respectively; DNA staining with DAPI (I); and merged (J) and (K) (shown in [L]).

(M) Expression of ET-HMGN1. Lanes 1–3 as in (F). Lanes 4–7, extracts of two pairs, each matching the same patient (1 or 2), of normal breast (No) and ductal invasive carcinoma G2 (Ca) tissue samples. Lane 8, ovarian tumor. Lanes 9 and 10, normal colon and colon adenocarcinoma G2.

speckles serve as depots of mRNA splicing factors and supply the needed factors at the demand of the translation machinery [13]. Even so, with some reports showing that the splicing speckles could themselves operate as transcription sites [14], their function remains obscure. A number of nuclear proteins are guided to the speckles through the arginine-and-serine-rich domain [15]. It is possible that a built-in message in the ET sequence also directs ET-H1.0 toward the splicing speckles.

We also performed western blot analyses with the anti-ET-HMGN1 antibodies. This revealed that ET-HMGN1 is present in the nuclear fraction of MCF 7 cells and, like ET-H1.0, occurs in all analyzed normal and cancerous tissues (Figure 3M). We estimated the relative abundance of ET-HMGN1 compared to HMGN1 to be about 1:50 (Appendix S6).

Abundance of ET-H1.0 Generally Does Not Correlate with H1.0

Having established the existence and cellular localization of the ET proteins, we then searched for their possible regulation in biological processes. Butyrate is known to reduce the growth of many cell types, causing an arrest in the G1 phase of the cell cycle [16]. In HeLa S3 cells, the synthesis of H1.0 occurs at a very low level (mRNA and protein) but is dramatically augmented by butyrate treatment [17]. We found that in the butyrate-stimulated cells, high expression levels of H1.0 were accompanied by substantial increase of ET-H1.0. The increase was detectable on western blots (Figures 4A and 4B) and by immunofluorescence staining (Figures 4E and 4F).

Given that butyrate has been shown to induce apoptosis in human cancer cells [18], we decided to analyze the occurrence of H1.0 and its ET-H1.0 forms upon apoptosis induced by DTT [19], Vinblastine [20], and TNF-α [21]. Unlike the case of butyrate stimulation, in the cells treated with DTT, Vinblastine, and TNF-α, only ET-H1.0 was augmented. The strongest induction of ET-H1.0 was observed after treatment of the cells with DTT or TNF-α for 30 hr. In apoptotic cells, ET-H1.0 was localized in the shrunken and vacuolarized nuclei within the condensed chromatin, but the ET-H1.0-containing speckles were also observed outside the nuclei. These results suggest that the levels of ET-H1.0 are regulated independently of translation of *h1.0* gene. The nature of this regulation mechanism, as well as a potential role of ET-H1.0 in differentiation and apoptosis, remains to be elucidated. Given that the N-terminal-extension region of ET-H1.0 is highly basic (pI > 12), we speculate that it could increase the tightness of histone binding to DNA. In this way, the extension could lead to formation of more

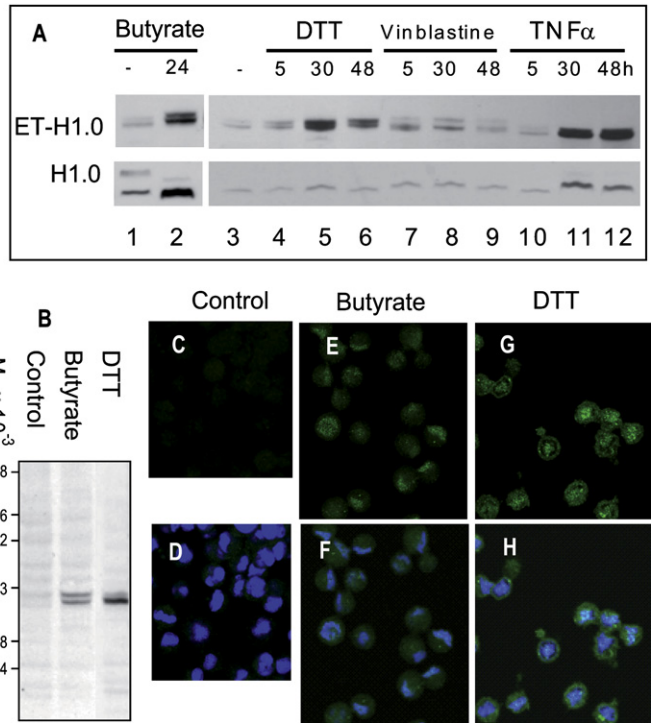


Figure 4. In HeLa S3 Cells, ET-H1.0 Can Be Stimulated by Butyrate and Other Agents, Causing Apoptosis

(A) Western blot analysis of perchloric acid extracts of HeLa S3 cells treated with butyrate, DTT, Vinblastine, and TNFα. The blots were probed with anti-ET-H1.0 and H1.0 antibodies.

(B) Western blot analysis of whole lysates of the untreated (lane 1), butyrate-treated, and DTT-treated cells (24 hr).

(C–H) Immunofluorescence analysis of cells treated with butyrate and DTT. (C), (E), and (G): anti-ET-H1.0 staining. (D), (F), and (H): merged antibody and DAPI staining.

compacted chromatin that is characteristic of both terminally differentiated and apoptotic cells.

Conclusions

Our work presents strong evidence that some 5′-noncoding sequences of human genes can be translated into proteins upon creation of the new start codon during or after transcription. The N-terminally extended proteins occur in normal human cells, and their synthesis is not related to alterations at the DNA level. For the examples described here, the amounts of ET proteins appear to be about two orders of magnitude lower than those of the standard products of the corresponding genes. However, given that H1.0 and HMGN1 occur in about 10^6–10^7 copies per cell [22–24], the ET forms are still relatively abundant compared to transcription factors and other chromatin proteins with specific regulatory functions. Presumably, the ET proteins were not identified until now because they are not predictable from the corresponding genes and occur "in the shadow" of their normal forms. Thus, in the past, the ET proteins were probably observed in many experiments but simply ignored as artifacts. The discovery of two ET proteins implies the existence of a dedicated enzymatic machinery of RNA editing, which is probably used more widely. Modern high-resolution MS data may already contain many more examples of unexpected protein forms, which could be brought to light by in-depth informatics analysis.

Experimental Procedures

The experimental procedures are described in the Supplemental Data file, available online.

Supplemental Data

Supplemental Data include Supplemental Experimental Procedures and six appendices and can be found with this article online at http://www.current-biology.com/S0960-9822(08)01294-3.

Acknowledgments

We thank Sonja Krüger for technical assistance. We are grateful to Yong Zhang and Jürgen Cox for expert bioinformatic advice. This work was supported by the Max Planck Society for the Advancement of Science, HEROIC, a 6th Framework grant of the European Commission, and the Munich Center for Integrated Protein Science (CIPSM).

Received: June 27, 2008
Revised: August 28, 2008
Accepted: September 19, 2008
Published online: November 6, 2008

References

1. Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198–207.
2. Mann, M., Hendrickson, R.C., and Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. Annu. Rev. Biochem. *70*, 437–473.
3. Cravatt, B.F., Simon, G.M., and Yates, J.R., 3rd. (2007). The biological impact of mass-spectrometry-based proteomics. Nature *450*, 991–1000.
4. Kuster, B., Schirle, M., Mallick, P., and Aebersold, R. (2005). Scoring proteomes with proteotypic peptide probes. Nat. Rev. Mol. Cell Biol. *6*, 577–583.
5. Bustin, M. (1999). Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins. Mol. Cell. Biol. *19*, 5237–5246.
6. Ding, H.F., Bustin, M., and Hansen, U. (1997). Alleviation of histone H1-mediated transcriptional repression and chromatin compaction by the acidic activation region in chromosomal protein HMG-14. Mol. Cell. Biol. *17*, 5843–5855.
7. West, K.L. (2004). HMGN proteins play roles in DNA repair and gene expression in mammalian cells. Biochem. Soc. Trans. *32*, 918–919.
8. Maresca, T.J., and Heald, R. (2006). The long and the short of it: linker histone H1 is required for metaphase chromosome compaction. Cell Cycle *5*, 589–591.
9. Wiśniewski, J.R., Zougman, A., Kruger, S., and Mann, M. (2007). Mass spectrometric mapping of linker histone H1 variants reveals multiple acetylations, methylations, and phosphorylation as well as differences between cell culture and tissue. Mol. Cell. Proteomics *6*, 72–87.
10. Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. (1993). dbEST–database for "expressed sequence tags". Nat. Genet. *4*, 332–333.
11. Gott, J.M., and Emeson, R.B. (2000). Functions and mechanisms of RNA editing. Annu. Rev. Genet. *34*, 499–531.
12. Ochsenreiter, T., Anderson, S., Wood, Z.A., and Hajduk, Z.L. (2008). Alternative RNA Editing Produces a Novel Protein Involved in Mitochondrial DNA Maintenance in Trypanosomes. Mol. Cell. Biol., in press.
13. Lamond, A.I., and Spector, D.L. (2003). Nuclear speckles: a model for nuclear organelles. Nat. Rev. Mol. Cell Biol. *4*, 605–612.
14. Moen, P.T., Jr., Johnson, C.V., Byron, M., Shopland, L.S., de la Serna, I.L., Imbalzano, A.N., and Lawrence, J.B. (2004). Repositioning of muscle-specific genes relative to the periphery of SC-35 domains during skeletal myogenesis. Mol. Biol. Cell *15*, 197–206.
15. Caceres, J.F., Misteli, T., Screaton, G.R., Spector, D.L., and Krainer, A.R. (1997). Role of the modular domains of SR proteins in subnuclear localization and alternative splicing specificity. J. Cell Biol. *138*, 225–238.
16. Prasad, K.N., and Sinha, P.K. (1976). Effect of sodium butyrate on mammalian cells in culture: a review. In Vitro *12*, 125–132.

17. D'Anna, J.A., Gurley, L.R., and Tobey, R.A. (1983). Extent of histone modifications and H1(0) content during cell cycle progression in the presence of butyrate. Exp. Cell Res. *147*, 407–417.
18. Chopin, V., Toillon, R.A., Jouy, N., and Le Bourhis, X. (2002). Sodium butyrate induces P53-independent, Fas-mediated apoptosis in MCF-7 human breast cancer cells. Br. J. Pharmacol. *135*, 79–86.
19. Tartier, L., McCarey, Y.L., Biaglow, J.E., Kochevar, I.E., and Held, K.D. (2000). Apoptosis induced by dithiothreitol in HL-60 cells shows early activation of caspase 3 and is independent of mitochondria. Cell Death Differ. *7*, 1002–1010.
20. Upreti, M., Lyle, C.S., Skaug, B., Du, L., and Chambers, T.C. (2006). Vinblastine-induced apoptosis is mediated by discrete alterations in subcellular location, oligomeric structure, and activation status of specific Bcl-2 family members. J. Biol. Chem. *281*, 15941–15950.
21. Wajant, H., Pfizenmaier, K., and Scheurich, P. (2003). Tumor necrosis factor signaling. Cell Death Differ. *10*, 45–65.
22. Kuehl, L., Salmond, B., and Tran, L. (1984). Concentrations of high-mobility-group proteins in the nucleus and cytoplasm of several rat tissues. J. Cell Biol. *99*, 648–654.
23. Crippa, M.P., Pash, J.M., Gerwin, B.I., Smithgall, T.E., Glazer, R.I., and Bustin, M. (1990). Expression of chromosomal proteins HMG-14 and HMG-17 in transformed human cells. Cancer Res. *50*, 2022–2026.
24. Zlatanova, J., and Doenecke, D. (1994). Histone H1 zero: a major player in cell differentiation? FASEB J. *8*, 1260–1268.
25. Steen, H., and Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. *5*, 699–711.

Supplemental Data

Evidence for Insertional RNA Editing in Humans

Alexandre Zougman, Piotr Ziółkowski, Matthias Mann, and Jacek R. Wiśniewski

Supplemental Experimental Procedures

Human tissue

Human tissue was retrieved during surgery. Informed consent was obtained, and the study was approved by the local ethics committee (Medical Academy of Wroclaw, Poland).

Cell culture

All cells were grown under 5% CO_2 at 37°C. MCF7, HeLa and A549 cells were grown in DMEM medium (GIBCO) supplemented with 10% fetal bovine serum (FBS). HeLa S3 cells were grown in suspension in RPMI 1640 medium (GIBCO) supplemented with 10% FBS. Exponentially growing HeLa S3 cells were subjected to treatment with 5 mM butyrate for 24 hours, 5 mM DTT for 5, 30 and 48 hours, 80 nM Vinblastine for 5, 30 and 48 hours, and 150 ng/ml TNF-α for 5, 30 and 48 hours.

Protein extraction from human tissue and cultured cells

The entire protein extraction procedure was carried out at 4°C in a cold room. 0.4-2 g of frozen tissue was cut into small pieces and extracted with 3 vol. (m/v) of 5 % (v/v) $HClO_4$ using an IKA Ultra Turbax blender at maximum speed of approximately 25,000 rpm for 20-30 s and centrifuged at 15,000 $\times$ g for 5 min. Fat was removed from the top of the tubes with a spatula, the supernatant was collected and the pellet was re-extracted with 3 vol. of 5 % (v/v) $HClO_4$. Proteins were precipitated from the combined supernatants with 33% (w/v) CCl_3COOH for 30 min and collected by centrifugation at 15,000 $\times$ g for 10 min. Culture cells were extracted with 5 % (v/v) $HClO_4$ using three freezing/thawing cycles as described previously [1].

LC-MS/MS analysis

In-gel protein tryptic digestion was performed essentially as previously described [2]. The resulting peptide mixtures were desalted using in-house made C_{18} STAGE tips [3], vacuum-dried and reconstituted in 0.5% acetic acid prior to analysis.

Peptide mixtures were separated by online reversed-phase nanoscale capillary liquid chromatography and analyzed by electrospray tandem mass spectrometry. The LC-MS/MS setup was similar to that described before [4]. Samples were injected onto an in-house made 15 cm reversed phase spraying fused-silica capillary column (inner diameter 75 μm, packed with 3 μm ReproSil-Pur C18-AQ media (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany), using either Agilent 1100 nanoflow (Agilent Technologies, Palo Alto, CA) or Proxeon EASY-nLC (Proxeon Biosystems, Odense, Denmark) systems. The LC setup was connected to either LTQ Orbitrap or LTQ-FT mass spectrometers (Thermo Electron, Bremen, Germany) equipped with a nanoelectrospray ion source (Proxeon Biosystems, Odense, Denmark). The peptides were separated with 100 or 200 min gradients from 5 to 40% CH_3CN in 0.5% acetic acid. Data-dependent acquisition was employed. Survey MS scans were acquired either in the orbitrap with the resolution set to a value of 60,000 or in FT cell with the resolution set to a value of 100,000. Up to the five most intense ions per scan were fragmented and analyzed in the linear trap. Additionally, high accuracy analysis of fragmentation was also performed in the orbitrap with the resolution set to 15,000.The lock-mass option was used as previously described for increased accuracy in mass measurements [5]. The spectra were searched either against an NCBI Human EST database or in-house created database containing the sequences of ET- proteins with the use of MASCOT (Matrix Science, London, UK) search engine [6] followed by manual verification.

Antibodies

Antibodies against ET extension of H1.0 and HMGN1 were elicited in rabbits using RAGPLEGAFNSRTR peptide (residues -17 **–** -4) and LPRRHPARAFPAA (residues -14**–** -2),

respectively. The peptides were conjugated to ovalbumin (Imject Maleimide Activated Ova, Pierce). Animals were injected with 0.2 mg of crosslinked protein. Mono-specific antibodies were purified on affinity columns which were prepared by coupling of individual peptides to iodoacetate activated gel (SulfoLink, Pierce) according to the manufacturer's protocol. Antibodies against H1.0, HMGN, and SC-35 were purchased from Biozol (mouse monoclonal ab 11079, Eching, Germany), ABCAM (rabbit polyclonal ab 5212, Cambridge, UK), and Novus Biologicals (mouse monoclonal, NB100-1774, Littleton, USA), respectively.

Western blot analysis

For Western analysis, proteins were transferred from SDS gels onto the nitrocellulose membrane by electroblotting at 10 V/cm for 40 min. The proteins were crosslinked to the membrane by incubation in 0.5% (v/v) glutaraldehyde in PBS for 10 min. The membranes were blocked with 10% (v/v) normal goat serum (NGS) for 30 min prior to incubation with the primary antibodies together with 1% NGS in PBS containing 0.1% Tween 20 (PBS-T). Following 2 h incubation, the membrane was extensively washed with PBS-T. Primary antibodies were visualized on the membrane with ECL peroxidase-conjugated IgGs.

Immunofluorescence

Cells were fixed in methanol at -20°C for 3 min and nonspecific binding was blocked by incubation in 10% (v/v) normal goat serum. The affinity purified primary polyclonal antibodies against the ET-extension of H1.0 were used at concentration of 2 μg/ml. The monoclonal anti-H1.0 was used at concentration of 25 μg/ml. The primary antibodies were visualized with secondary AlexaFluor488 or 546 goat antibodies (Molecular Probes) at concentration of 2.6 μg/ml. Nuclei were counterstained with DAPI at concentration of 1 μg/ml. Analyses were performed using LEICA TCS SP2 Confocal Laser Scanning Microscope.

References

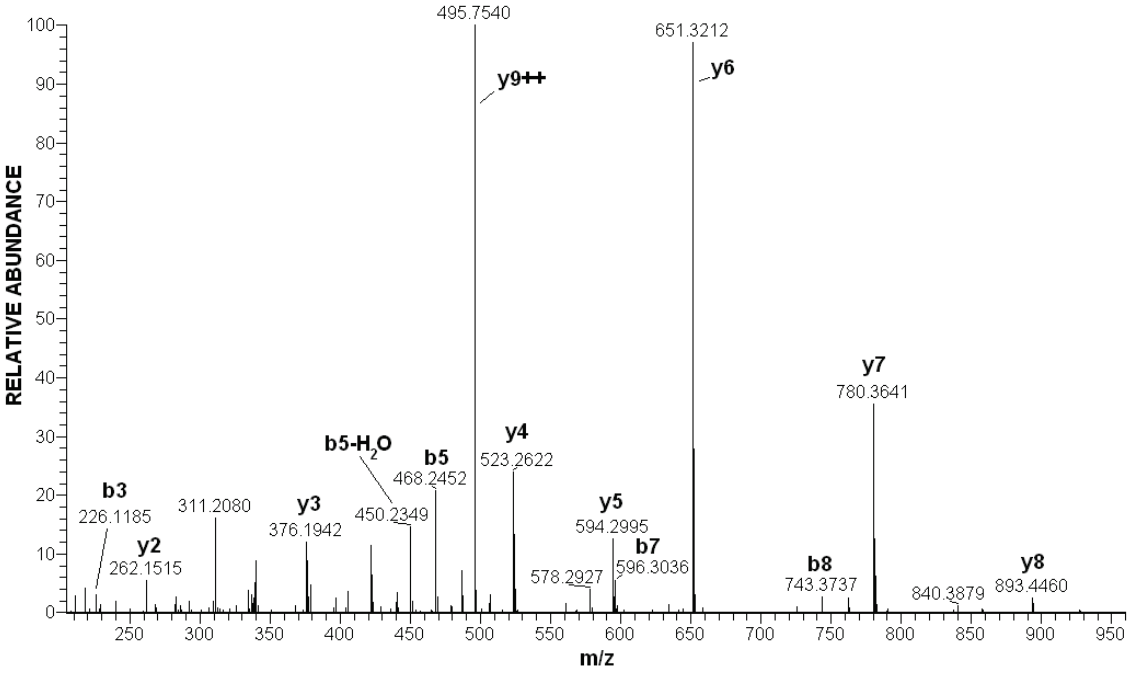
1. Zougman, A., and Wiśniewski, J.R. (2006). Beyond linker histones and high mobility group proteins: global profiling of perchloric acid soluble proteins. J. Proteome Res. *5*, 925–934.
2. Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996). Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. Anal. Chem. *68*, 850–858.
3. Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal. Chem. *75*, 663–670.
4. Olsen, J.V., and Mann, M. (2004). Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. Proc. Natl. Acad. Sci. USA *101*, 13417–13422.
5. Olsen, J.V., de Godoy, L.M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. (2005). Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. Mol. Cell. Proteomics *4*, 2010–2021.
6. Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551–3567.

Appendix S1. MS/MS Spectra of the Intrinsic ET-H1.0 Peptides

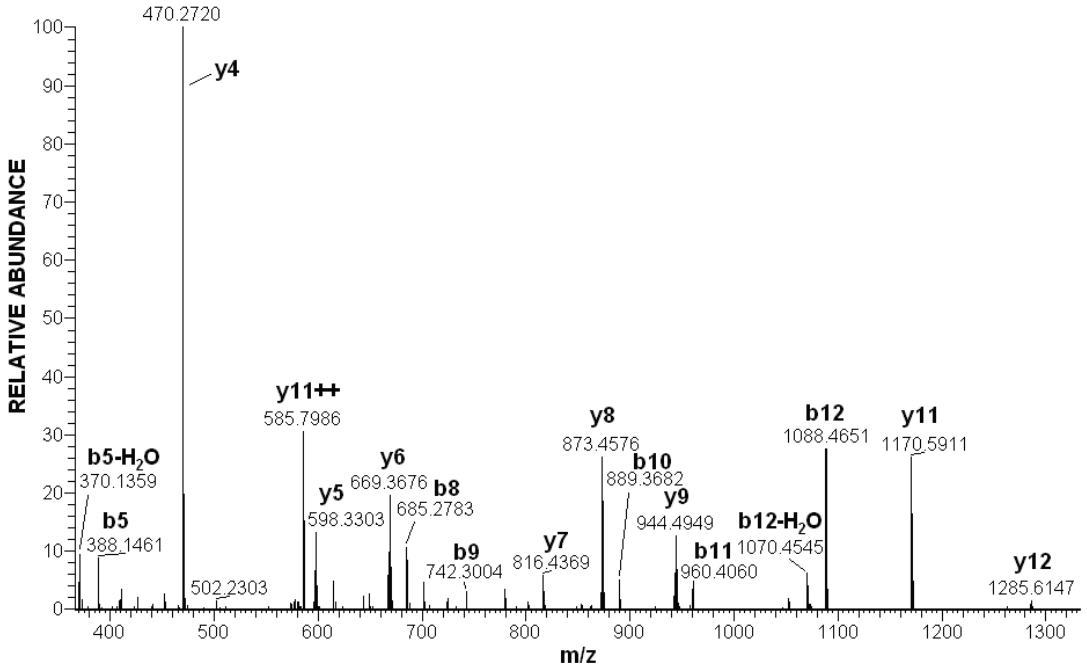
Spectra for the depicted ET-H1.0 peptides were obtained by fragmenting the peptides in the linear ion trap section of a hybrid linear ion trap orbitap instrument (LTQ-Orbitrap). Fragments were transferred to the orbitrap and measured with high resolution (15,000) and with low parts per million mass accuracy. Together with the excellent representation of canonical *b* and *y* ions, this establishes the identity of the peptides with virtually complete certainty.

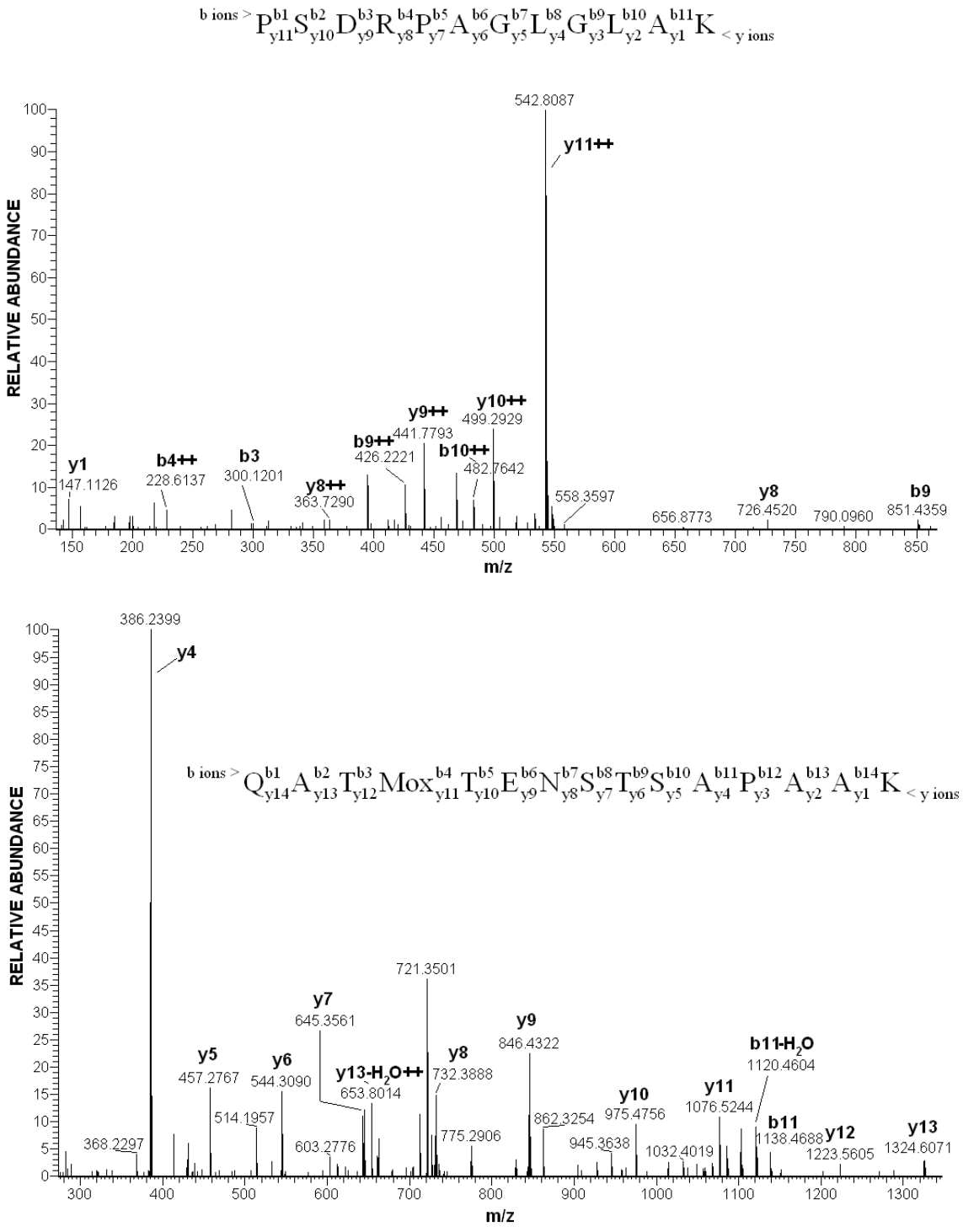
Peptide	M/Z	Charge	Mass observed, Da	Mass calculated, Da	Delta mass, Da
AGPLEGAFNSR	559.783005	2	1117.551458	1117.551559	-0.000101
GSGADPEAGFAQPPTR	779.368752	2	1556.722952	1556.721878	0.001074
PSDRPAGLGLAK	394.559129	3	1180.655559	1180.656372	-0.000813
QATMoxTENSTSAPAAK	762.351853	2	1522.689154	1522.693329	-0.004175
QATMoxTENSTSAPAAKPK	583.620661	3	1747.840155	1747.841049	-0.000894
QRQSphosPVPR	524.258236	2	1046.50192	1046.502213	0.000293
RALPTPEPGR	547.308827	2	1092.603102	1092.603958	-0.000856
SSDSSLASPGAALQTGPVVR	950.493481	2	1898.97241	1898.969711	0.002699

b ions > $A^{b1}_{y10}G^{b2}_{y9}P^{b3}_{y8}L^{b4}_{y7}E^{b5}_{y6}G^{b6}_{y5}A^{b7}_{y4}F^{b8}_{y3}N^{b9}_{y2}S^{b10}_{y1}R$ < y ions

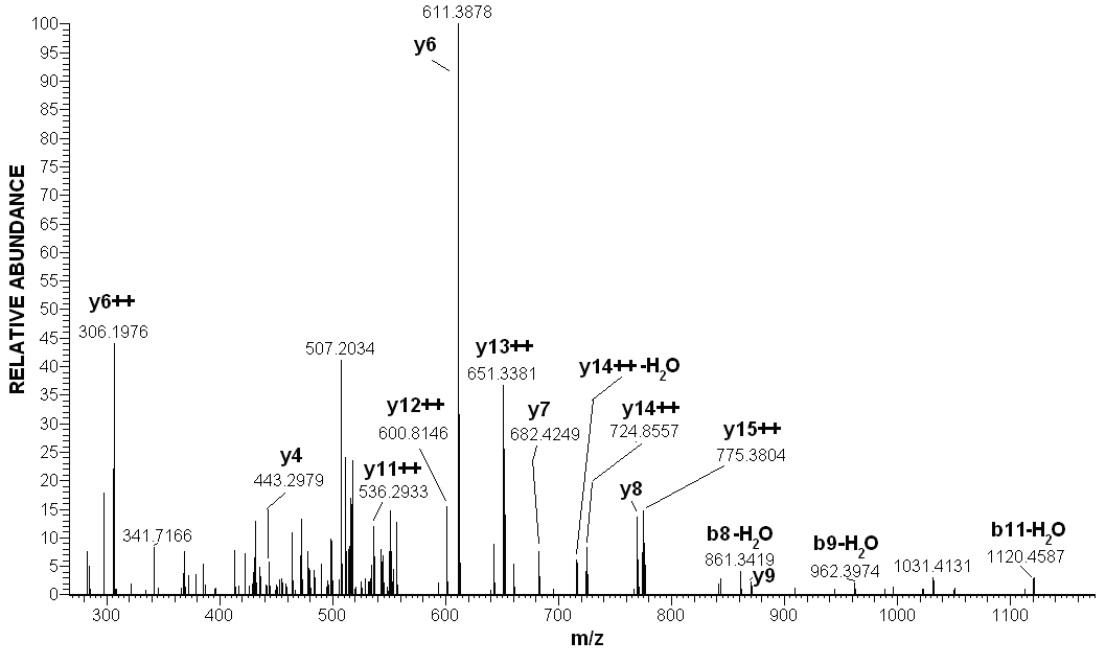


b ions > $G^{b1}_{y15}S^{b2}_{y14}G^{b3}_{y13}A^{b4}_{y12}D^{b5}_{y11}P^{b6}_{y10}E^{b7}_{y9}A^{b8}_{y8}G^{b9}_{y7}F^{b10}_{y6}A^{b11}_{y5}Q^{b12}_{y4}P^{b13}_{y3}P^{b14}_{y2}T^{b15}_{y1}R$ < y ions

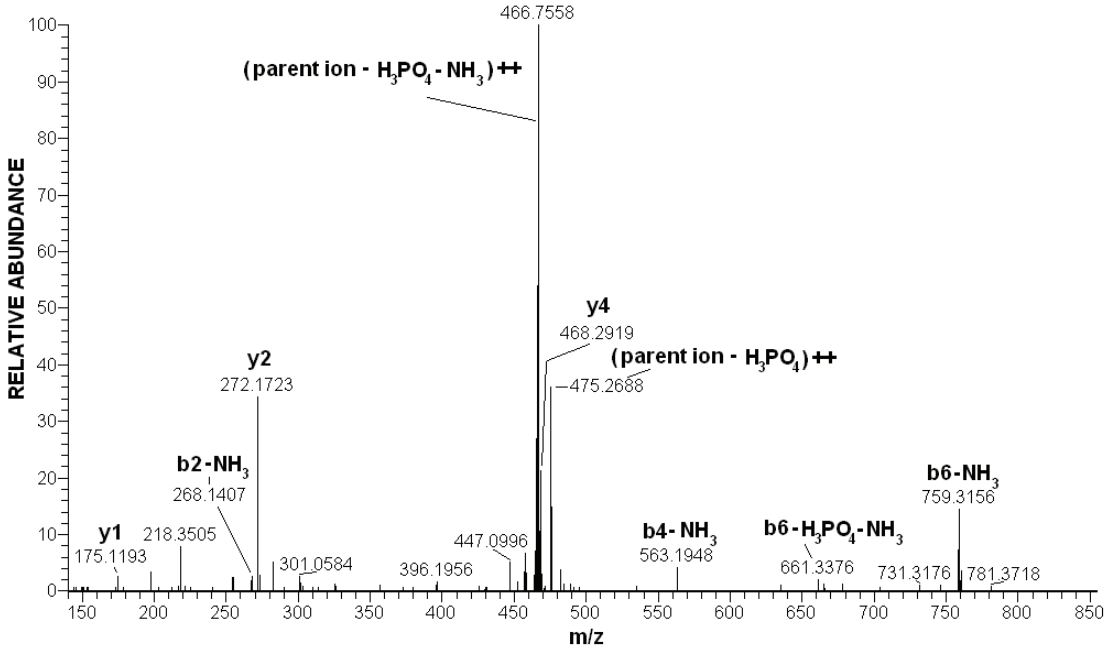


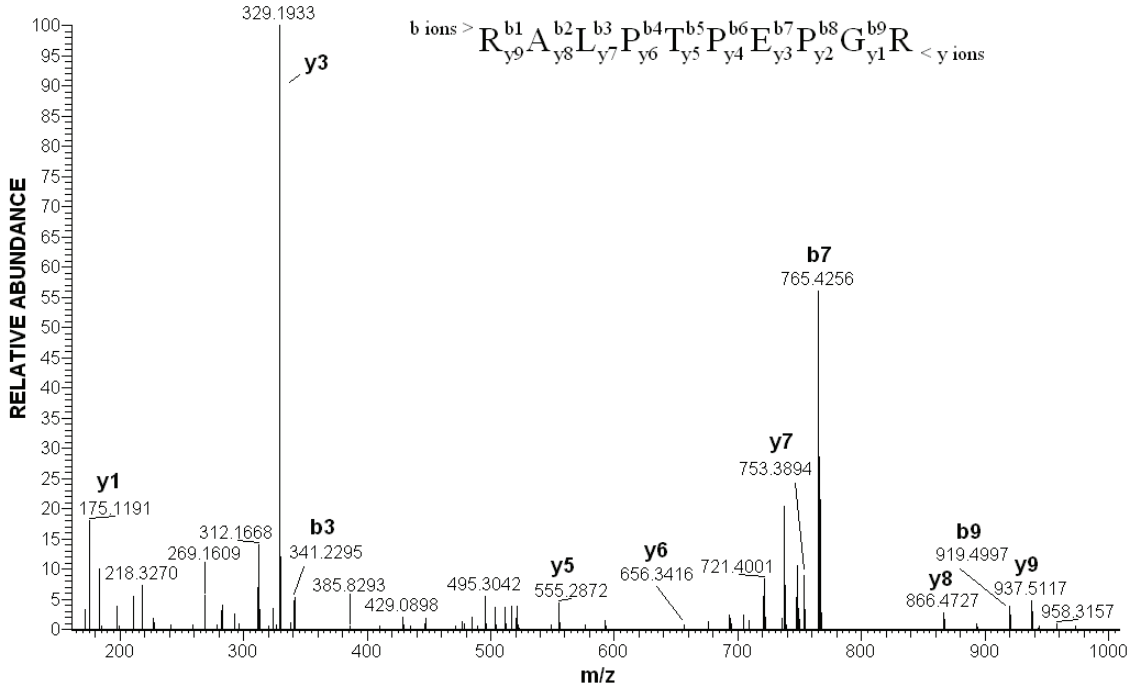

b ions > P y11 S y10 D y9 R y8 P y7 A y6 G y5 L y4 G y3 L y2 A y1 K < y ions
RELATIVE ABUNDANCE
m/z
542.8087
y11++
y10++
499.2929
y9++
441.7793
b9++
426.2221
b10++
482.7642
y1
147.1126
b4++
228.6137
b3
300.1201
y8++
363.7290
558.3597
656.8773
y8
726.4520
790.0960
b9
851.4359
386.2399
y4
b ions > Q y14 A y13 T y12 Mox y11 T y10 E y9 N y8 S y7 T y6 S y5 A y4 P y3 A y2 A y1 K < y ions
721.3501
y7
645.3561
y9
846.4322
y5
457.2767
y6
544.3090
y13-H2O++
653.8014
y8
732.3888
b11-H2O
1120.4604
y10
975.4756
y11
1076.5244
514.1957
862.3254
b11
1138.4688
y12
1223.5605
y13
1324.6071
368.2297
603.2776
775.2906
945.3638
1032.4019

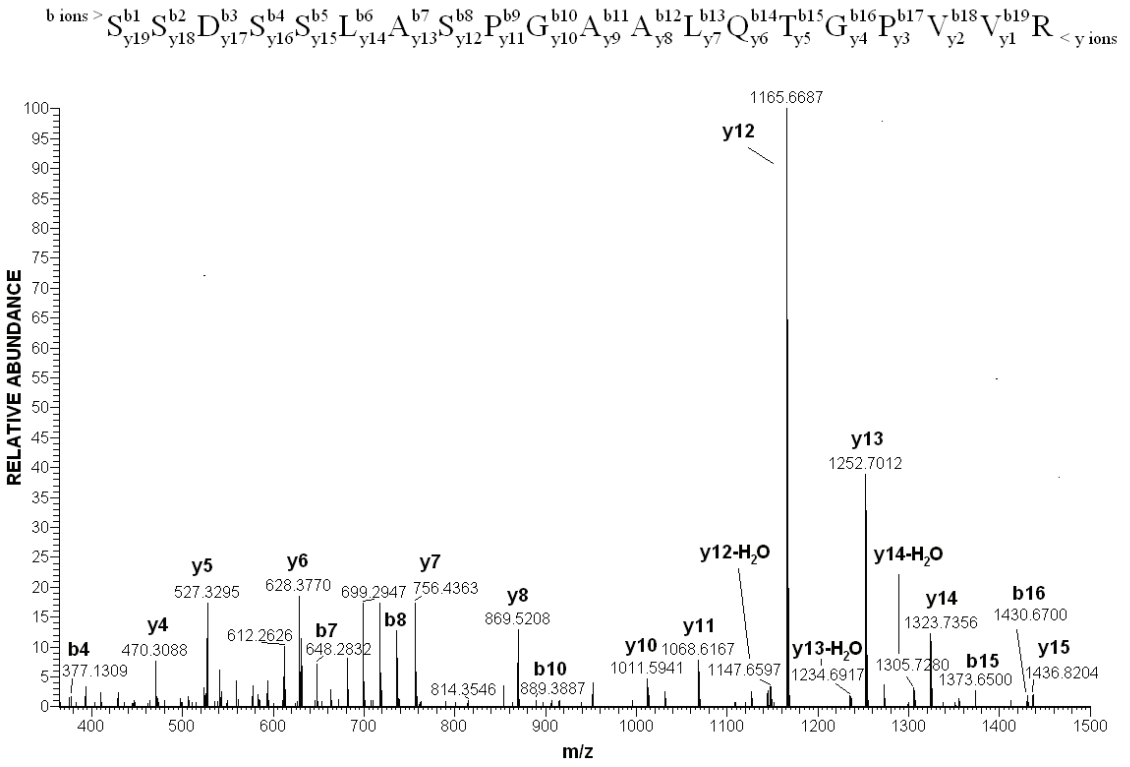
b ions > $Q^{b1}_{y16}A^{b2}_{y15}T^{b3}_{y14}Mox^{b4}_{y13}T^{b5}_{y12}E^{b6}_{y11}N^{b7}_{y10}S^{b8}_{y9}T^{b9}_{y8}S^{b10}_{y7}A^{b11}_{y6}P^{b12}_{y5}A^{b13}_{y4}A^{b14}_{y3}K^{b15}_{y2}P^{b16}_{y1}K$ < y ions



b ions > $Q^{b1}_{y7}R^{b2}_{y6}Q^{b3}_{y5}Sphos^{b4}_{y4}P^{b5}_{y3}V^{b6}_{y2}P^{b7}_{y1}R$ < y ions

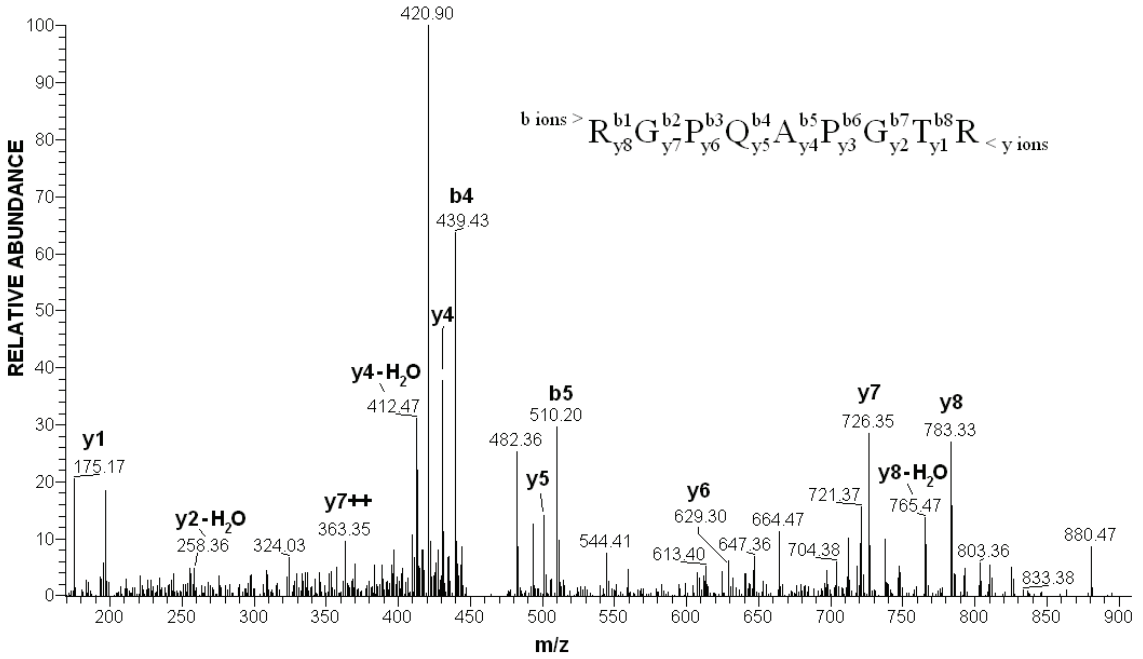


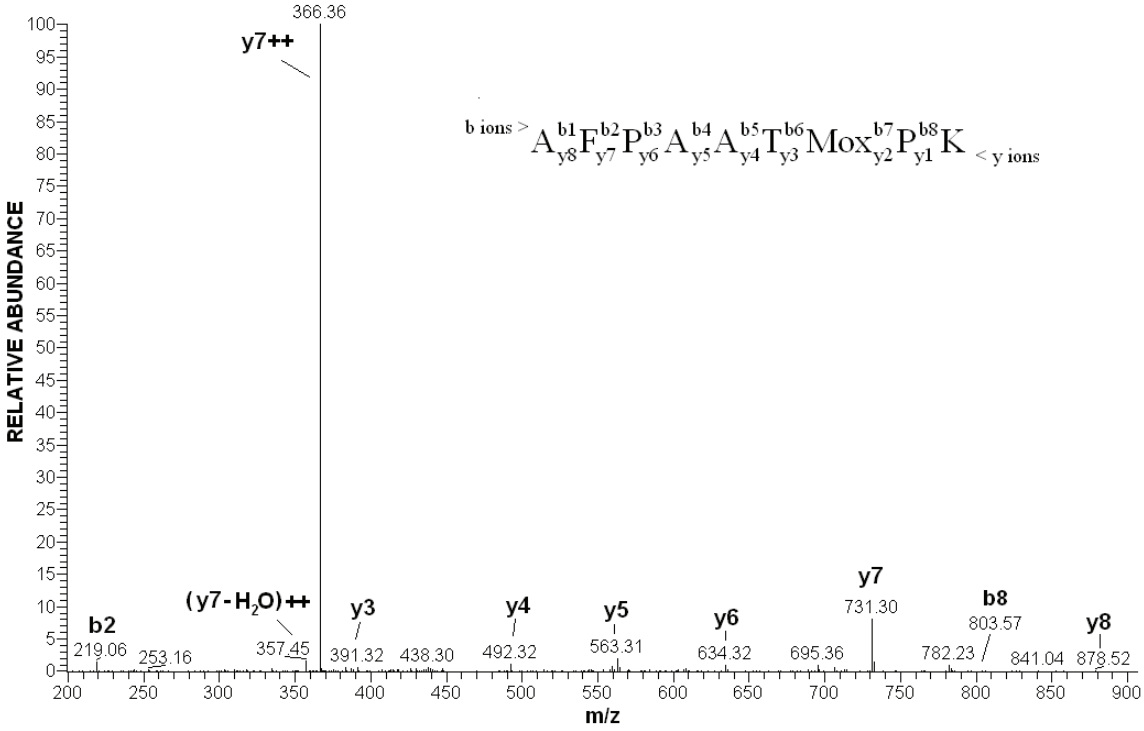

b ions > $R^{b1}_{y9}A^{b2}_{y8}L^{b3}_{y7}P^{b4}_{y6}T^{b5}_{y5}P^{b6}_{y4}E^{b7}_{y3}P^{b8}_{y2}G^{b9}_{y1}R$ < y ions
RELATIVE ABUNDANCE
m/z
329.1933
y3
b7
765.4256
y7
753.3894
y1
175.1191
312.1668
269.1609
218.3270
b3
341.2295
385.8293
429.0898
495.3042
y5
555.2872
y6
656.3416
721.4001
b9
919.4997
y8
866.4727
y9
937.5117
958.3157


b ions > $S^{b1}_{y19}S^{b2}_{y18}D^{b3}_{y17}S^{b4}_{y16}S^{b5}_{y15}L^{b6}_{y14}A^{b7}_{y13}S^{b8}_{y12}P^{b9}_{y11}G^{b10}_{y10}A^{b11}_{y9}A^{b12}_{y8}L^{b13}_{y7}Q^{b14}_{y6}T^{b15}_{y5}G^{b16}_{y4}P^{b17}_{y3}V^{b18}_{y2}V^{b19}_{y1}R$ < y ions
RELATIVE ABUNDANCE
m/z
1165.6687
y12
y13
1252.7012
y12-H2O
y14-H2O
y14
1323.7356
b16
1430.6700
y15
1436.8204
y5
527.3295
y6
628.3770
y7
699.2947
756.4363
y8
869.5208
y4
470.3088
612.2626
b7
648.2832
b8
b4
377.1309
814.3546
b10
889.3887
y10
1011.5941
y11
1068.6167
1147.6597
y13-H2O
1234.6917
1305.7280
b15
1373.6500

Appendix S2. MS/MS Spectra of the Intrinsic ET-HMGN1 Peptides

Peptide	M/Z	Charge	Mass observed, Da	Mass calculated, Da	Delta mass, Da
AFPAATMoxPK	475.242370	2	948.470188	948.473877	-0.003689
RGPQAPGTR	470.258391	2	938.502230	938.504578	-0.002348


RELATIVE ABUNDANCE
m/z
b ions > $R_{y8}^{b1}G_{y7}^{b2}P_{y6}^{b3}Q_{y5}^{b4}A_{y4}^{b5}P_{y3}^{b6}G_{y2}^{b7}T_{y1}^{b8}R$ < y ions
420.90
b4
439.43
y4
y4-H2O
412.47
b5
510.20
482.36
y5
y1
175.17
y2-H2O
258.36
324.03
y7++
363.35
544.41
y6
629.30
613.40
647.36
664.47
704.38
721.37
y7
726.35
y8
783.33
y8-H2O
765.47
803.36
833.38
880.47


RELATIVE ABUNDANCE
m/z
b ions > $A_{y8}^{b1}F_{y7}^{b2}P_{y6}^{b3}A_{y5}^{b4}A_{y4}^{b5}T_{y3}^{b6}Mox_{y2}^{b7}P_{y1}^{b8}K$ < y ions
366.36
y7++
(y7-H2O)++
357.45
b2
219.06
253.16
y3
391.32
438.30
y4
492.32
y5
563.31
y6
634.32
695.36
y7
731.30
b8
803.57
782.23
841.04
y8
878.52

Appendix S3. Human EST Sequences Pertaining to ET-H1.0 and ET-HMGN1

Human EST Sequences Pertaining to ET-H1.0

GenBank accession #	Origin
BX328722	Placenta
BX371207	Jurkat cell line
BX386531	Jurkat cell line
BX397992	Placenta
BX410654	Fetal brain
BX418456	Fetal brain
BX432449	Fetal brain
BX432450	Fetal brain
BX410653	Fetal brain
BX426875	Fetal liver
BX419072	Fetal brain

A Human EST Sequence Pertaining to ET-HMGN1

GenBank accession #	Origin
CV807291	Blastocyst

Appendix S4. Multiple Sequence Alignment of ET-H1.0 ESTs Carrying the Insertion

Only the ET part of the alignment is shown.

```
BX397992        -GGATGCTGGGAAAAGGGAGGCAGAGGAGGCGGAGGCAGAGGCAGAGGCAGAGCCCGGTG 59
BX432450        -GGATGCTGGGAAAAGGGAGGCAGAGGAGGCGGAGGCAGAGGCAGAGGCAGAGCCCGGTG 59
BX418456        -GGATGCTGGGAAAAGGGAGGCAGAGGAGGCGGAGGCAGAGGCAGAGGCAGAGCCCGGTG 59
BX371207        -GGATGCTGGGAAAAGGGAGGCAGAGGAGGCGGAGGCAGAGGCAGAGGCAGAGCCCGGTG 59
BX419072        GGGATGCTGGGAAAMGGGAGGCAGAGGAGGCGGAGGCAGAGGCAGAGGCAGAGCCCGGTG 60
BX386531        -GGATGCTGGGAAAAGGGAGGCAGAGGAGGCGGAGGCAGAGGCAGAGGCAGAGCCCGGTG 59
BX432449        -GGATGCTGGGAAAAGGGAGGCAGAGGAGGCGGAGGCAGAGGCAGAGGCAGAGCCCGGTG 59
BX426875        GGGATGCTGGGAAAARGGAGGCAGAGGAGGCGGAGGCAGAGGCAGAGGCAGAGCCCGGTG 60
BX328722        -GGATGCTGGGAAAAGGGAGGCAGAGGAGGCGGAGGCAGAGGCAGAGGCAGAGCCCGGTG 59
BX410653        -GGATGCTGGGAAAAGGGANGCAGAGGAGGCGGAGGCAGAGGCAGAGGCAGAGCCCGGTG 59
BX410654        -GGATGCTGGGAAAAGGGAGGCAGAGGAGGCGGAGGCAGAGGCAGAGGCAGAGCCCGGTG 59
                 *************  *** ****************************************

BX397992        CCGAGACCAAGCGACAGACCGGCGGGGCTGGGCCTCGCAAAGCCGGCTCGGCGAGCTCTC 119
BX432450        CCGAGACCAAGCGACAGACCGGCGGGGCTGGGCCTCGCAAAGCCGGCTCGGCGAGCTCTC 119
BX418456        CCGAGACCAAGCGACAGACCGGCGGGGCTGGGCCTCGCAAAGCCGGCTCGGCGAGCTCTC 119
BX371207        CCGAGACCAAGCGACAGACCGGCGGGGCTGGGCCTCGCAAAGCCGGCTCGGCGAGCTCTC 119
BX419072        CCGAGACCAAGCGACAGACCGGCGGGGCTGGGCCTCGCAAAGCCGGCTCGGCGAGCTCTC 120
BX386531        CCGAGACCAAGCGACAGACCGGCGGGGCTGGGCCTCGCAAAGCCGGCTCGGCGAGCTCTC 119
BX432449        CCGAGACCAAGCGACAGACCGGCGGGGCTGGGCCTCGCAAAGCCGGCTCGGCGAGCTCTC 119
BX426875        CCGAGACCAAGCGACAGACCGGCGGGGCTGGGCCTCGCAAAGCCGGCTCGGCGAGCTCTC 120
BX328722        CCGAGACCAAGCGACAGACCGGCGGGGCTGGGCCTCGCAAAGCCGGCTCGGCGAGCTCTC 119
BX410653        CCGAGACCAAGCGACAGACCGGCGGGGCTGGGCCTCGCAAAGCCGGCTCGGCGAGCTCTC 119
BX410654        CCGAGACCAAGCGACAGACCGGCGGGGCTGGGCCTCGCAAAGCCGGCTCGGCGAGCTCTC 119
                ************************************************************

BX397992        CCGACACCCGAGCCGGGGAGGAAAAGCAGCGACTCCTCGCTCGCATCCCCGGGAGCCGCA 179
BX432450        CCGACACCCGAGCCGGGGAGGAAAAGCAGCGACTCCTCGCTCGCATCCCCGGGAGCCGCA 179
BX418456        CCGACACCCGAGCCGGGGAGGAAAAGCAGCGACTCCTCGCTCGCCTCCCCGGGAGCCGCC 179
BX371207        CCGACACCCGAGCCGGGGAGGAAAAGCAGCGACTCCTCGCTCGCATCCCCGGGAGCCGCA 179
BX419072        CCGACACCCGAGCCGGGGAGGAAAAGCAGCGACTCCTCGCTCGCATCCCCGGGAGCCGCA 180
BX386531        CCGACACCCGAGCCGGGGAGGAAAAGCAGCGACTCCTCGCTCGCATCCCCGGGAGCCGCA 179
BX432449        CCGACACCCGAGCCGGGGAGGAAAAGCAGCGACTCCTCGCTCGCATCCCCGGGAGCCGCA 179
BX426875        CCGACACCCGAGCCGGGGAGGAAAAGCAGCGACTCCTCGCTCGCATCCCCGGGAGCCGCA 180
BX328722        CCGACACCCGAGCCGGGGAGGAAAAGCAGCGACTCCTCGCTCGCATCCCCGGGAGCCGCA 179
BX410653        CCGACACCCGAGCCGGGGAGGAAAAGCAGCGACTCCTCGCTCGCATCCCCGGGAGCCGCA 179
BX410654        CCGACACCCGAGCCGGGGAGGAAAAGCAGCGACTCCTCGCTCGCATCCCCGGGAGCCGCA 179
                ******************************************* **************

BX397992        CTCCAGACTGGCCCGGTAGTCAGGGGCTCAGGAGCAGAWCCCGAGGCAGGCTTTGCTCAG 239
BX432450        CTCCAGACTGGCCCGGTAGTCAGGGGCTCAGGAGCAGATCCCGAGGCAGGCTTTGCTCAG 239
BX418456        CTCCAGACTGGCCCGGTAGTCAGGGGCTCAGGAGCAGATCCCGAGGCAGGCTTTGCTCAG 239
BX371207        CTCCAGACTGGCCCGGTAGTCAGGGGCTCAGGAGCAGATCCCGAGGCAGGCTTTGCTCAG 239
BX419072        CTCCAGACTGGCCCGGTAGTCAGGGGCTCAGGAGCAGATCCCGAGGCAGGCTTTGCTCAG 240
BX386531        CTCCAGACTGGCCCGGTAGTCAGGGGCTCAGGAGCAGATCCCGAGGCAGGCTTTGCTCAG 239
BX432449        CTCCAGACTGGCCCGGTAGTCAGGGGCTCAGGAGCAGATCCCGAGGCAGGCTTTGCTCAG 239
BX426875        CTCCAGACTGGCCCGGTAGTCAGGGGCTCAGGAGCAGATCCCGAGGCAGGCTTTGCTCAG 240
BX328722        CTCCAGACTGGCCCGGTAGTCAGGGGCTCAGGAGCAGATCCCGAGGCAGGCTTTGCTCAG 239
BX410653        CTCCAGACTGGCCCGGTAGTCAGGGGCTCAGGAGCAGATCCCGAGGCAGGCTTTGCTCAG 239
BX410654        CTCCAGACTGGCCCGGTAGTCAGGGGCTCAGGAGCAGATCCCGAGGCAGGCTTTGCTCAG 239
                ************************************ ********************

BX397992        CCTCCGACGAGGGCTGGCCCTTTGGAAGGCGCCTTCAACAGCCGGACCAGACAGGCCACC 299
BX432450        CCTCCGACGAGGGCTGGCCCTTTGGAAGGCGCCTTCAACAGCCGGACCAGACAGGCCACC 299
BX418456        CCTCCGACGAGGGCTGGCCCTTTGGAAGGCGCCTTCAACAGCCGGACCAGACAGGCCACC 299
BX371207        CCTCCGACGAGGGCTGGCCCTTTGGAAGGCGCCTTCAACAGCCGGACCAGACAGGCCACC 299
BX419072        CCTCCGACGAGGGCTGGCCCTTTGGAAGGCGCCTTCAACAGCCGGACCAAACAGGCCACC 300
BX386531        CCTCCGACGAGGGCTGGCCCTTTGGAAGGCGCCTTCAACAGCCGGACCAGACAGGCCACC 299
BX432449        CCTCCGACGAGGGCTGGCCCTTTGGAAGGCGCCTTCAACAGCCGGACCAGACAGGCCACC 299
BX426875        CCTCCGACGAGGGCTGGCCCTTTGGAAGGCGCCTTCAACAGCCGGACCAGACAGGCCACC 300
BX328722        CCTCCGACGAGGGCTGGCCCTTTGGAAGGCGCCTTCAACAGCCGGACCAGACAGGCCACC 299
BX410653        CCTCCGACGAGGGCTGGCCCTTTGGAAGGCGCCTTCAACAGCCGGACCAGACAGGCCACC 299
BX410654        CCTCCGACGAGGGCTGGCCCTTTGGAAGGCGCCTTCAACAGCCGGACCAGACAGGCCACC 299
                ********************************************** **********
```

Appendix S5. Translation of the Nucleotide Sequences of the ET Extensions

Predicted ET part of H1.0

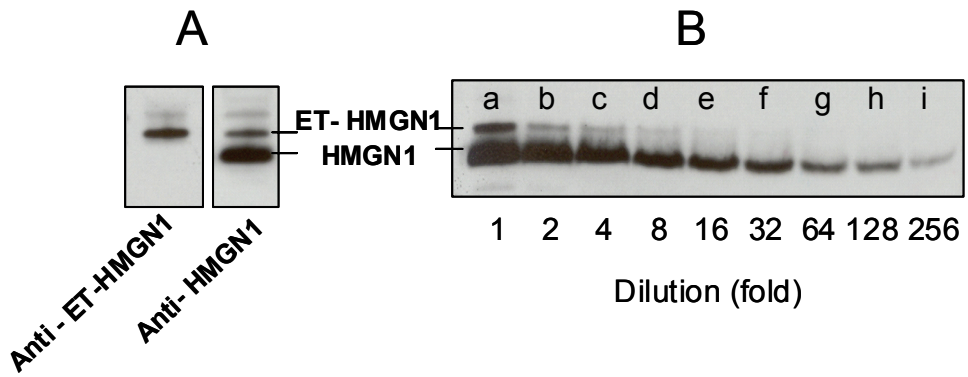
```
ggatgctgggaaaagggaggcagaggaggcggaggcagaggcagaggcagagcccggtgccg
   M  L  G  K  G  R  Q  R  R  R  R  Q  R  Q  R  Q  S  P  V  P
agaccaagcgacagaccggcggggctgggcctcgcaaagccggctcggcgagctctcccg
 R  P  S  D  R  P  A  G  L  G  L  A  K  P  A  R  R  A  L  P
acacccgagccggggaggaaaagcagcgactcctcgctcgcatccccgggagccgcactc
 T  P  E  P  G  R  K  S  S  D  S  S  L  A  S  P  G  A  A  L
cagactggcccggtagtcaggggctcaggagcagatcccgaggcaggctttgctcagcct
 Q  T  G  P  V  V  R  G  S  G  A  D  P  E  A  G  F  A  Q  P
ccgacgagggctggccctttggaaggcgccttcaacagccggaccagacaggccaccatg…
 P  T  R  A  G  P  L  E  G  A  F  N  S  R  T  R  Q  A  T  M
```

Predicted ET part of HMGN1

```
…gggatgctcgggcggcgggaggagtggcagcggcaaggcagcccagtttcgcgaaggctg
  G  M  L  G  R  R  E  E  W  Q  R  Q  G  S  P  V  S  R  R  L
tcggcgcgccgcggcccgcaggcacccggcacgcgccttccccgcaggcacccggcacgc
 S  A  R  R  G  P  Q  A  P  G  T  R  L  P  R  R  H  P  A  R
Gccttccccgccgccacgatg…
 A  F  P  A  A  T  M
```

Appendix S6. Estimation of the ET-HMGN1 to HMGN1 Ratio in MCF7 Cells

Since the commercially available antibodies against HMGN1 recognize both "normal" and ET forms of the proteins (below, panel A), we were able to compare the levels of both protein forms on western blots. Series of dilutions of perchloric extracts from MCF7 cells were western blotted and the staining intensities of HMGN1 and ET-HMGN1 were compared (below, panel B). The intensity of ET-HMGN1band in the undiluted sample (lane a) was similar to intensities of HMGN1 bands in 32-64 fold diluted sample (lanes f and g). Thus, we conclude that in MCF7 cells ET-HMGN1 occurs at levels of about 2-3% of HMGN1.



A, Western blots of perchloric acid extracts of MCF7 cells probed with antibodies against ET-HMGN1 and HMGN1, respectively. **B,** Dilution series of the protein extracts probed with anti-HMGN1 antibodies.

Future directions

It is my interest to further probe the function of the extended forms of nuclear proteins. The discovery of two ET-proteins implies the existence of a dedicated enzymatic machinery of RNA editing, which is probably used more widely. I would like to find possible specific binding partners of the ET part of H1.0 as well as to perform RNA pull-down experiments in order to identify potential components of the suggested RNA-editing machinery. Modern high resolution MS data may already contain many more examples of unexpected protein forms, which will surely be brought to light by thorough analysis.

CSF proteome and peptidome profiling

Cerebrospinal fluid (CSF) is a liquid produced primarily by the choroid plexus epithelial cells. It provides brain and central nervous system (CNS) with nutrients, stimuli and protection. The blood-CSF barrier at the choroid plexus is formed between the epithelial cells which are linked by tight junctions and fenestrated blood capillaries. Cerebrospinal fluid continuously circulates through the ventricles, over the surface of the brain, and is absorbed at the arachnoid villi and at the cranial and spinal nerve root sheaths. CSF is produced at the rate of about 0.3 ml per min. Total CSF volume varies from 100 to 150 ml and its turnover rate is approximately 6 hours (*75*, *76*). For diagnostic purposes CSF is usually aspirated by lumbar puncture. CSF protein concentration is one of the most sensitive indicators of pathology within the CNS. Newborn patients have up to 150 mg per dL (1.5 g per L) of protein. The adult range of 18 to 58 mg per dL (0.18 to 0.58 g per L) is reached between six and 12 months of age (*77*).

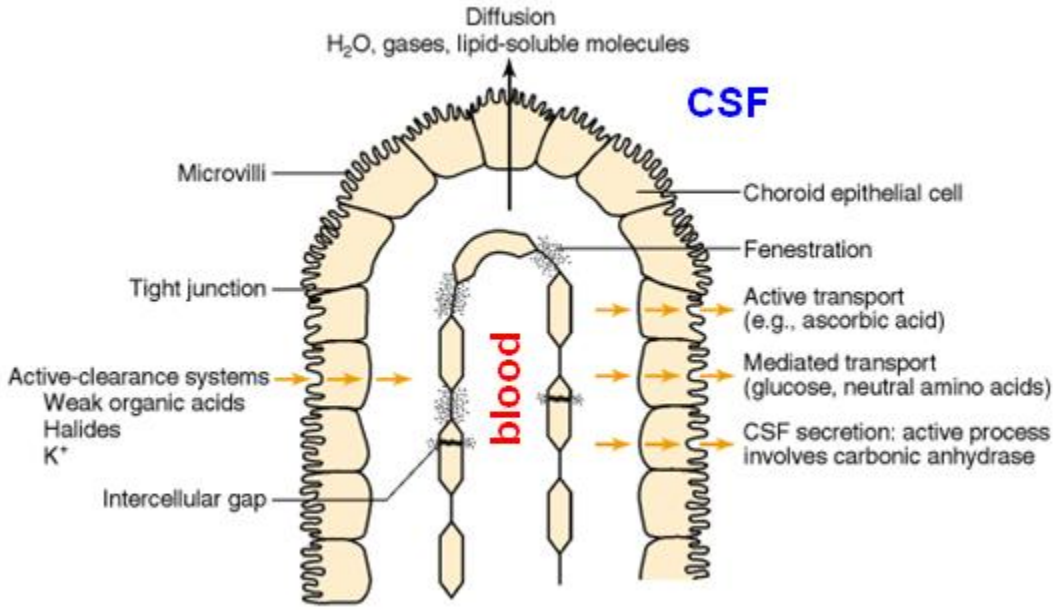


Figure 27. The blood—CSF barrier. Choroid plexus epithelial cells are joined by tight junctions. Microvilli are present on the CSF-facing surface of the choroid plexus cells increasing the surface area and aiding in fluid secretion. The molecules can permeate the barrier either passively or actively (from ref. (*76*)).

Similar to blood, the prevalence of a small number of highly abundant proteins such as albumin, tranferrin or immunoglobulins makes the identification of the low abundant proteins in CSF very challenging. It is important to stress that CSF contains not only proteins but also neuropeptides which are shuttled by the CSF to their final destination targets. As a direct recipient of cell-shedding products, it is a potential indicator of abnormal CNS states such as inflammation, infection, neurodegenerative processes, and tumor growth. Until recently, little was known about CSF polypeptide content. The rapid improvement of proteomics technologies presented researchers with opportunities to initiate CSF proteome profiling studies (*78-82*). However, the number of publications pertaining to the CSF peptidome (i.e. endogenous peptides as opposed to proteins) remains very low. Our strategy was to perform a *combined* analysis of both proteome and peptidome CSF content of clinically normal patients and to create a depository of cerebrospinal fluid peptide and protein identifications so they can serve as a reference for future research. With our publication we also wanted to give an impulse to neural peptidome discovery studies. We used high accuracy and resolution mass spectrometry in both MS and MS/MS modes for characterization of the CSF peptidome content. Endogenous peptides are often produced by non-typical enzyme cleavage. Thus, the database searches has to be performed with a "no enzyme specificity" option and the software processing time was significantly slower compared with the search when the enzyme cleavage site is known. Additionally, current software algorithms are not optimized for the "no enzyme specificity" search and do not fully utilize the high accuracy MS/MS data content provided by the advanced mass spectrometry instrumentation. Thus, we had to manually re-check the peptide identifications and *de-novo* sequence some of the post-translationally modified neuropeptides in order to modify the offset parameters of the search engine. Figure 28 presents some of the identified post-translationally modified peptides, the glycopeptides.

Protein	Peptide	Modification
Probable G-protein coupled receptor 37 precursor	EGWTIALPGRA	O-linked HexNAc1Hex1NeuAc1
Probable G-protein coupled receptor 37 precursor	HEGWTIALPGRA	O-linked HexNAc1Hex1NeuAc1
Probable G-protein coupled receptor 37 precursor	HEGWTIALPGRA	O-linked HexNAc1Hex1NeuAc2
Heparin-binding EGF-like growth factor precursor	GLAAGTSNPDPPTVSTDQLLPLGGGRDRKV	O-linked HexNAc1Hex1NeuAc1
Fractalkine precursor	LGVLITPVPDAQAATRRQAV	Amide (C-term) O-linked HexNAc1Hex1NeuAc2
Fractalkine precursor	LGVLITPVPDAQAATRRQ	O-linked HexNAc1Hex1NeuAc1
Fractalkine precursor	LGVLITPVPDAQAATRRQAVG	O-linked HexNAc1Hex1NeuAc1
Fractalkine precursor	LGVLITPVPDAQAATRRQA	O-linked HexNAc1Hex1NeuAc2
Fractalkine precursor	LGVLITPVPDAQAATRRQAVG	O-linked HexNAc1Hex1NeuAc2
Fractalkine precursor	LGVLITPVPDAQAATRRQAVGL	O-linked HexNAc1Hex1NeuAc2
Fractalkine precursor	LGVLITPVPDAQAATRRQAVGLLA	O-linked HexNAc1Hex1NeuAc2
Fractalkine precursor	LGVLITPVPDAQAATRRQAVGLLAF	O-linked HexNAc1Hex1NeuAc2
Fractalkine precursor	LGVLITPVPDAQAATRRQAVGLLAFLG	O-linked HexNAc1Hex1NeuAc2
Alpha-2-HS-glycoprotein precursor	TVVQPSVGAAAGPVVPPCPGRIRHFKV	O-linked HexNAc1Hex1NeuAc1
Esophageal cancer related gene 4 protein	EAPVPTKTKVAVDENKAKEFLGSLKRQ	O-linked HexNAc1Hex1NeuAc1
Esophageal cancer related gene 4 protein	EAPVPTKTKVAVDENKAKEFLGSLKRQ	O-linked HexNAc1Hex1NeuAc2
SEL-OB protein	DDFLDTVQETATSIGNAKSSRI	O-linked HexNAc1Hex1NeuAc2
Hypothetical protein FLJ26517	AAVGTSAAPVPSDNH	O-linked HexNAc1Hex1NeuAc2
Cadherin 20, type 2 preproprotein	DTPTPQGELEALLSDKPQSHQRT	O-linked HexNAc1Hex1NeuAc1
Cadherin 20, type 2 preproprotein	DTPTPQGELEALLSDKPQSHQ	O-linked HexNAc1Hex1NeuAc2
Cadherin 20, type 2 preproprotein	DTPTPQGELEALLSDKPQSHQRT	O-linked HexNAc1Hex1NeuAc2
Kexin type 1 inhibitor precursor	LETPAPQVPARRLLPP	O-linked HexNAc1Hex1NeuAc1
Kexin type 1 inhibitor precursor	AADHDVGSELPPEGVLGALLRV	O-linked HexNAc1Hex1NeuAc2
Isoform 1 of Insulin-like growth factor II precursor	DVSTPPTVLPDNFPRYPV	Amide (C-term) O-linked HexNAc1Hex1NeuAc2
Isoform 1 of Insulin-like growth factor II precursor	DVSTPPTVLPDNFPRYPVGKF	O-linked HexNAc1Hex1NeuAc1
Isoform 1 of Insulin-like growth factor II precursor	DVSTPPTVLPDNFPRYP	O-linked HexNAc1Hex1NeuAc2
Isoform 1 of Insulin-like growth factor II precursor	DVSTPPTVLPDNFPRYPVG	O-linked HexNAc1Hex1NeuAc2
Isoform 1 of Insulin-like growth factor II precursor	DVSTPPTVLPDNFPRYPVGKF	O-linked HexNAc1Hex1NeuAc2

Figure 28. CSF glycopeptides identified by high accuracy MS/MS in this study. With the *de-novo* sequencing identification of the *DVSTPPTVLPDNFPRYPVGKF* glycopeptide forms of IGF-II, the characteristic mass offsets of the discovered glycosylation signatures were taken into account while searching the protein database with MASCOT engine.

We combined high accuracy and high resolution MS with low resolution MS/MS acquisition for proteome profiling. As the result of our work, 563 peptide forms and about 800 proteins were identified by high stringency criteria. This was the first integrated proteome/peptidome content analysis of any body fluid reported in the literature. The study pointed out the potential of such analysis for CSF profiling and biomarker discovery. We showed that the major contributor to the protein population of CSF is protein secretion combined with ongoing proteolytic processes involved in cell surface remodeling and protein shedding. We also found, unexpectedly, that the CSF protein population differs in composition from that of plasma presumably due to functions of the CSF that are as yet mostly unexplored. This work also led to optimization of protocols for neuropeptide profiling and kindled our interest in identification and characterization of novel neuropeptides.

research articles Journal of proteome research

Integrated Analysis of the Cerebrospinal Fluid Peptidome and Proteome

Alexandre Zougman,[†] Bartosz Pilch,[†,‡] Alexandre Podtelejnikov,[§] Michael Kiehntopf,[△] Claudia Schnabel,[⊥] Chanchal Kumar,[†] and Matthias Mann*[,†,‡]

Department of Proteomics and Signal Transduction, Max-Planck-Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany, Center for Experimental BioInformatics (CEBI), Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark, Proxeon Bioinformatics A/S, Staermosegaardsvej 6, DK-5230 Odense M, Denmark, Institute of Clinical Chemistry and Laboratory Medicine, University of Jena, Germany, and Department of Clinical Chemistry, Center of Clinical Pathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Received August 03, 2007

Cerebrospinal fluid (CSF) is the only body fluid in direct contact with the brain and thus is a potential source of biomarkers. Furthermore, CSF serves as a medium of endocrine signaling and contains a multitude of regulatory peptides. A combined study of the peptidome and proteome of CSF or any other body fluid has not been reported previously. We report confident identification in CSF of 563 peptide products derived from 91 precursor proteins as well as a high confidence CSF proteome of 798 proteins. For the CSF peptidome, we use high accuracy mass spectrometry (MS) for MS and MS/MS modes, allowing unambiguous identification of neuropeptides. Combination of the peptidome and proteome data suggests that enzymatic processing of membrane proteins causes release of their extracellular parts into CSF. The CSF proteome has only partial overlap with the plasma proteome, thus it is produced locally rather than deriving from plasma. Our work offers insights into CSF composition and origin.

Keywords: cerebrospinal fluid • neuropeptides • LC-MS/MS • LTQ-Orbitrap • peptidomics • proteomics

Introduction

Cerebrospinal fluid (CSF) is produced mainly by the choroid plexus in the lateral, third, and fourth ventricles of the brain at an approximate rate of 500 mL per day in healthy adults, the total volume of CSF being about 135 mL.[1,2] With a turnover time of about 6 h, CSF is in constant flow within the brain ventricles and subarachnoid space of the brain and spinal cord, providing buoyancy and protection. It carries nutrients for cells, removes products of their metabolism, and serves as a transport medium for hormones. It not only contains polypeptides which pass through the blood−brain barrier but also harbors peptides and proteins manufactured locally. As a direct recipient of cell-shedding products, it is a potential indicator of abnormal CNS states such as inflammation, infection, neurodegenerative processes and tumor growth.

The low total protein concentration—the CSF/serum ratio of protein concentration is about 0.004—as well as high amounts of albumin and immunoglobins (about 70%)[3] challenge the identification of low-abundant proteins in CSF by 2D electrophoresis and make liquid chromatography–mass spectrometry-based technologies[4] the method of choice for CSF proteome profiling. Additionally, the resolution of electrophoresis-based methods is low for polypeptides smaller than 10 kDa, which makes ultrafiltration or precipitation combined with mass spectrometry practically the only option for low molecular weight proteome or "peptidome" profiling.[5]

CSF has been the subject of several proteomic studies during recent years. Published reports concentrated on discovering potential CSF biomarkers in neurodegenerative diseases,[6–8] multiple sclerosis,[9] traumatic brain injury,[10,11] and aging[12] as well as on the analysis of a broad inventory of CSF proteins.[13–16]

In contrast to CSF proteome studies, the number of publications related to the CSF peptidome is scarce—despite the tremendous interest in finding novel ligands for orphan receptors, particularly the GPCRs, the treasure trove of the pharmaceutical industry. Stark et al.[17] were the first to use organic phase extraction and mass spectrometry-based profiling to identify a number of peptides in human CSF. Yuan and Desiderio[18] employed ultrafiltration and mass spectrometry in a proof-of-principle CSF peptidomics study. Not much has been added to this knowledge afterward, notwithstanding the ongoing dramatic improvements in mass spectrometric and separation technologies.

Here we present the first in-depth and high-confidence study of the CSF peptidome, consisting of 563 identified peptides derived from 91 protein precursors in cerebrospinal fluid. We

* To whom correspondence should be addressed. E-mail: mmann@biochem.mpg.de. Fax: + 49 (0) 89 8578 3209. Tel.: +49 (0) 89 8578 2557.
† Max-Planck-Institute for Biochemistry.
‡ University of Southern Denmark.
§ Proxeon Bioinformatics A/S.
△ University of Jena.
⊥ University Medical Center Hamburg-Eppendorf.

10.1021/pr070501k CCC: $40.75
Published on Web 12/04/2007

also report the detection of 798 proteins in a parallel proteome profiling study by one-dimensional gel enhanced liquid chromatography–mass spectrometry (GeLC−MS). Some of the identified neuropeptides were found to carry features typical of regulatory peptides such as evidence that they were produced by proconvertase cleavage, N-terminal pyroglutamination, C-terminal amidation, and high cysteine content. Furthermore, half of their precursor proteins, including the precursors of known neuropeptides, were not identified in the proteome mapping experiment, suggesting that they are functional and not mere degradation products of abundant CSF proteins. Our high-confidence CSF proteome indicates that secreted and membrane bound proteins are by far the largest contributors. Interestingly, the CSF proteome is enriched in the membrane-bound receptor tyrosine phosphatase (PTPR) family, including the potential remyelination marker PTPRZ.

Our study was designed to give an impulse to the nascent CSF peptidome field and also to create a depository of cerebrospinal fluid peptides and proteins so they can serve as a reference for future research.

Experimental Details

Sample Collection and SDS-PAGE. CSF samples were taken by lumbar puncture for diagnostic purposes that turned out to be normal by standard clinical chemistry analysis. The study was approved by the ethical committee of the chamber of physicians in Hamburg and by the local ethics committee of the University of Jena. The samples were obtained by a pneumatic tube approximately 30–60 min after the CSF tap. Then aliquots were taken for clinical chemistry, and the CSF cytology was performed. The specimens were barcoded and frozen immediately after clinical laboratory analysis at −80 °C. These are the usual conditions for CSF analysis in a routine clinical setting. The total protein concentration, albumin, IgG, IgA, and IgM concentrations, the CSF/serum albumin ratio, and the IgG/albumin index revealed no signs of inflammation, blood−brain barrier leakage, or intrathecal IgG synthesis.[19,20] The CSF cell counts were within the reference range.[21] These samples are representative of a typical patient without CNS inflammation, e.g., for a "diagnostically normal individual".

Additional assessment of the five samples in the pooled fraction for the presence of hemoglobin by a human hemoglobin ELISA (Immundiagnostik AG; Bensheim, Germany) revealed no signs of blood contamination. The samples were consecutively reduced with dithiothreitol and alkylated by iodoacetamide before ultrafiltration. To collect the low molecular weight range fractions and to concentrate proteins, the samples were subjected to ultrafiltration on Centricon filter devices with a cutoff of 10 kDa (Millipore, Bedford, MA). Protein portions from the CSF samples of five individuals (0.5 mL of CSF each) were pooled; a protein portion of a sample from one individual, originating from 2 mL of CSF, was analyzed separately. Six ultrafiltrate low molecular weight fractions from the above-mentioned individuals were also analyzed separately. Proteins were resolved on 4–12% gradient NuPAGE Novex Bis−Tris gels (Invitrogen, Carlsbad, CA). The lanes were cut into 11 or 10 pieces and subjected to a standard in-gel trypsin digestion protocol. Briefly, the pieces were washed twice with 25 mM ammonium bicarbonate (ABC)/50% ethanol and dehydrated with absolute ethanol. After that, 0.5 μg of trypsin (Promega, Madison, WI) solution in 25 mM ammonium bicarbonate was added, and the enzyme was allowed to digest overnight at 37 °C. The peptide mixtures were extracted with 50% acetonitrile and 2% trifluoroacetic acid (TFA), and the organic solvent was evaporated in a vacuum centrifuge. Samples were reconstituted with 1% TFA, and STAGE tip purification was performed as previously described.[22]

The ultrafiltrate low molecular weight fractions were dried down in a vacuum centrifuge to a final volume of about 40 μL. An equal volume of 2% TFA was added, and the STAGE tip purification was performed.

LC-MS/MS and Data Analysis. Peptide mixtures were separated by online reversed-phase nanoscale capillary liquid chromatography (LC) and analyzed by electrospray tandem mass spectrometry. The samples were injected onto an in-house made 15 cm reversed-phase fused silica capillary column emitter (inner diameter 75 μm, packed with 3 μm ReproSil-Pur C_{18}-AQ media (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany)), using an Agilent 1100 nanoflow system (Agilent Technologies, Palo Alto, CA). The LC setup was connected to either a linear quadrupole ion trap-orbitrap (LTQ-orbitrap) or a linear quadrupole ion trap-Fourier transform (LTQ-FT) mass spectrometer (Thermo Electron, Bremen, Germany) equipped with a nanoelectrospray ion source (Proxeon Biosystems, Odense, Denmark).

In the case of proteome mapping using the LTQ-Orbitrap, the mass spectrometer was operated in the data dependent mode. Survey full scan MS spectra (from m/z 300 to 1600) were acquired in the orbitrap with resolution $R = 60\,000$ at m/z 400 (after accumulation to a target value of 1 000 000). The five or ten most intense ions were sequentially isolated for fragmentation and detection in the linear ion trap using collisionally induced dissociation at a target value of 30 000. Target ions already selected for MS/MS were dynamically excluded for 45 s.

In the case of proteome mapping on the LTQ-FT, the mass spectrometer was operated in the data-dependent mode to automatically switch between MS, MS^2, and MS^3 acquisition. Survey spectra in an m/z range of 300–1575 were acquired by FT-ICR, and the three most intense ions in the m/z ratio range 450–1400 were sequentially chosen for accurate mass measurement by FT-ICR-selected ion monitoring (SIM). They were subsequently fragmented in the ion trap to obtain MS^2 spectra. The most intense ion in MS^2 spectra was selected for another round of collisionally induced dissociation to obtain MS^3 spectra. The other MS conditions were the same as described previously.

For peptidome profiling on the LTQ-Orbitrap, survey full scan mass spectra (from m/z 300 to 2000) were acquired in the orbitrap after accumulation to a target value of 1 000 000 with resolution $R = 60\,000$. The five most intense ions were fragmented either in the ion trap or in the C-trap (Higher-energy C-trap dissociation, HCD). The MS/MS products were measured in the orbitrap with the resolution value set to 15 000 (after accumulation to a target value of 100 000). For accurate parent mass measurements, the lock-mass option was used as previously described.[23] Target ions already selected for MS/MS were dynamically excluded for 30 s.

The acquired data were searched against an in-house curated International Protein Index human protein sequence database[24] (IPI 3.17) with an automated database searching program, MASCOT (Matrix Science, London).[25] For proteome profiling, spectra were searched allowing a maximum mass deviation of 10 ppm and 0.4 Da for MS and MS^n peaks, respectively. MS^3 spectra were automatically scored with MSQuant, an open source program developed in-house (http://msquant.sourceforge.net) which parses MASCOT peptide iden-

tifications and enables their manual and automated validation. For peptidome profiling, spectra were searched with a maximum mass deviation of 10 ppm and 0.01 Da for MS and MS^2 peaks, respectively. All relevant hits were manually verified according to the following rules: (1) to be accepted for manual validation, the peptides needed to be at least seven amino acids in length and have a minimum MASCOT score of 15 (note that the mass accuracies in MS and MS/MS modes were extremely high and that MASCOT does not increase its score for high mass accuracy in MS and MS/MS data); (2) the majority of the most intense MS/MS ions should belong either to the *y* and *b* products of the precursor peptide or have a relevant link to these products, e.g., loss of H_2O from an ion containing a serine residue; (3) presence of the b_2 and a_2 ions; (4) if present, the *y* ion originating from the fragmentation at the proline residue and the corresponding *b* ion should have increased intensity; (5) continuity of the ion series (if discontinuous, a special explanation is required, e.g., the presence of the proline residue). An introduction to the rules of MS/MS sequencing can be found in the tutorial by Steen and Mann.[26] Some of the glycopeptides were partially sequenced de novo. With their sugar signatures revealed, the corresponding δ masses of the modifications were used in subsequent MASCOT searches.

To prepare our protein list, the peptide identifications were subjected to very stringent filtering. Only peptides with seven amino acids or longer were accepted. All of them were required to score above 27, a score indicated by MASCOT to be significant ($p < 0.05$). Proteins identified with at least two peptides were accepted. For one peptide hit identifications, in the case of LTQ-FT analysis, corresponding MS^3 validated spectra were required for positive identification. We ran a search against an IPI decoy database containing both forward and reversed sequences to estimate the number of false positive identifications present in the sample. Applying the same stringent criteria as we did for the forward database, searches of the reversed sequences resulted in no positive hits, confirming the reliability of our identifications.

Bioinformatics Analysis. For bioinformatics analysis, the ProteinCenter software package (Proxeon Bioinformatics A/S, Odense, Denmark) was used. ProteinCenter integrates public sequence databases to form a comprehensive and consistent superset. Due to support of major protein sequence databases and computational enrichment, ProteinCenter decreases the redundancy of the databases and significantly improves protein annotation. Signal peptide and transmembrane region predictions were automatically calculated by the software package. Gene Ontology distributions were calculated based on the GO Slim categories (http://www.geneontology.org/GO.slims.html) that give a broad overview of the ontology content. Filtering on peptidase activity was performed according to enzyme nomenclature EC 3.4. For the fold change analysis of GO categories, we used BiNGO, a Cytoscape plug-in[27] to find statistically over-represented GO categories in our CSF proteome map, taking as a reference the whole human proteome. A custom GO annotation for the reference IPI human data set was created by extracting GO annotations available for Human IPI IDs from EBI GOA Human 39.0 release (28 873 protein annotations).[28] The analysis was done using the "HyperGeometric test", and all GO terms which were significant with $p < 0.001$, after correcting for multiple term testing by Benjamini & Hochberg False Discovery Rate corrections, were selected as overrepresented.[29]

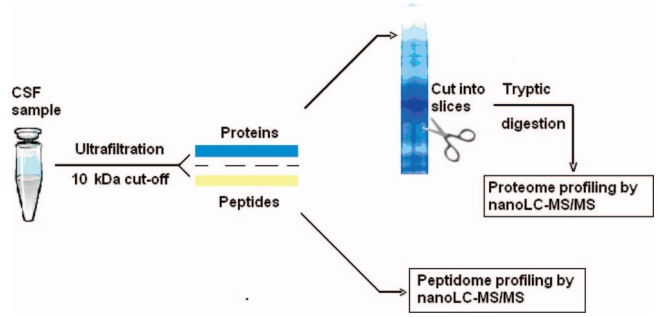


Figure 1. Overview of the procedure used for the analysis of the CSF peptidome and proteome.

Results and Discussion

The samples were obtained by lumbar puncture from six individuals undergoing diagnostic investigation. Typically 0.5 mL of CSF sample was available for analysis, which considering the low concentration of proteins in CSF made this investigation challenging in terms of sensitivity. By means of ultrafiltration, the CSF samples were split into low molecular "peptidome" and higher molecular "proteome" fractions (Figure 1). The peptidome fractions were analyzed directly, by capillary liquid chromatography and high accuracy mass spectrometric acquisition on LTQ-Orbitrap in both MS and MS/MS modes using different and complementary fragmentation modes.

For the CSF proteome study, we analyzed a 2 mL sample from one individual as well as the pooled higher molecular weight fraction of the 0.5 mL samples of five individuals. The large dynamic range of protein concentrations, where the overwhelming presence of a few proteins overshadows the rest, makes cerebrospinal fluid an analytic challenge similar to plasma. For example, the proportion of albumin and immunoglobulins can be as high as 65–70% of the total CSF protein content, creating difficulties in identification of the less-abundant proteins. As we have shown in previous studies, one-dimensional SDS-PAGE separation combined with tryptic digestion of the excised bands and subsequent LC-MS runs can successfully analyze the proteome to considerable depth.[30] We separated the CSF proteomes using 1D gels and analyzed them by LC MS/MS. The final CSF protein data set was derived by extensive protein profiling of the single CSF sample combined with the analysis of the pooled CSF samples. An overview of the experimental workflow is presented in Figure 1.

CSF Peptidome. In conventional proteome mapping experiments, proteins are digested with trypsin, and each protein is usually identified by at least two fully tryptic and nonmodified peptides. Tryptic peptides carry C-terminal basic amino acids protonated under acidic conditions facilitating their mass spectrometric fragmentation and consequent interpretation of the data. The identification of neuropeptides is much more challenging. They do not terminate in predictable amino acids, making the number of candidates to be considered in the database search much higher, and the nonstandard charge distribution within the peptide sequence can produce fragmentation spectra with a smaller number of characteristic ions. Most importantly, neuropeptides can be post-translationally modified, additionally complicating their analysis. To address these challenges, we employed a hybrid mass spectrometer—the LTQ-orbitrap—capable of high-resolution, high-accuracy analysis in the orbitrap part of the instrument and analyzed both the peptides and their fragmentation products in the

orbitrap. Additionally, we utilized a novel fragmentation method termed *higher-energy C-trap dissociation* (HCD) for MS/MS.[31] This allowed us to observe the full mass range of fragments in the spectra, facilitating the identification of peptides by adding the information of characteristic low mass ions such as y_1, y_2, a_2, b_2, and reporter ions of post-translational modifications. This contrasts with the normal mode of operation of ion traps in which fragment spectra are obtained at low resolution in the ion trap part of the instrument and with a low mass cutoff at about one-third of the precursor mass to charge (m/z) ratio.

In this way, from the six individuals, 563 peptides, including known neuropeptides, originating from 91 protein precursors were identified with extremely high stringency and manual verification (Supporting Information Table 1). The fact that several hundred neuropeptides can be identified at high stringency from a small volume clinical sample implies both high analytical sensitivity of our methods and rich peptide content of CSF. We first analyzed broad features of the CSF peptidome. A remarkable trait of the identified precursor proteins is that 78 of them (84%) are predicted to contain a signal peptide sequence—one of the features of secreted proteins in general and prepro-hormones in particular. A total of 46 precursor proteins (Supporting Information Table 2) did not appear in our CSF proteome data set, and some of these proteins are known precursors of neuropeptides. During a typical hormone processing event, the signal peptide is removed from a prepro-hormone by a signal peptidase.[32] The pro-hormone can be cleaved by specific convertases at dibasic sites, generating precursors which then are acted upon by carboxypeptidases[33] giving rise to mature forms of peptide hormones. In addition to this, peptides can become C-terminally amidated by peptidylglycine monooxigenase which requires a C-terminal glycine in the peptide sequence.[34] For some of the peptides, for example, the corticotropin-releasing hormone, the amidation is crucial for receptor binding.[35] In our data set, 23 of the identified peptides were produced by a typical dibasic cleavage—meaning a dibasic site N-terminal and C-terminal to the peptide or a signal peptide site N-terminal and a dibasic site C-terminal to the peptide. These included the proteolysis products of proSAAS and 7B2 proteins, the inhibitors of the most well-known pro-hormone convertases PC1 and PC2, respectively[36] (Supporting Information Table 3).

Pyroglutamic acid can be an essential element in functioning of biologically active peptides.[37] Hinkle et al.,[38] for example, demonstrated its importance in thyrotropin-releasing hormone receptor binding. Of the identified peptides, we found that 53 carry pyroglutamate at the N-terminus, which made this the most abundant modification in our data set. It should be stressed, though, that the majority of the pyroglutamate modifications of N-termini of peptides observed in proteomics experiments is introduced during the sample preparation process. Thus, it is impossible to suggest a functional meaning of a neuropeptide carrying solely N-terminal pyroglutamate without any other characteristic neuropeptide features.

Many biologically active peptides, such as insulin or endothelin, contain disulfide bridges which not only determine the functional structure of the molecules but also often protect them from proteolysis.[39,40] Additionally, some of the antimicrobial endogenous peptides, defensins, are disulfide bonded into specific structures.[41] We identified a set of cysteine-containing peptides, usually with an even number of cysteine residues. In some cases, peptides also contained an odd number of cysteine residues, the "spare" or sole cysteine residue functioning, presumably, in homodimerization.

The γ-secretase complex cleaves type I membrane proteins within their transmembrane domains.[42] Specifically, γ-secretase releases the amyloid β peptide (Aβ) that accumulates in the brains of patients with Alzheimer's disease.[43] Our data contain a number of possible secretase-related peptide products that in addition to being derived mostly from type I membrane proteins also reveal the characteristic secretase intramembrane cleavage signature at their C-termini (Supporting Information Table 4). Some of the peptides were found to be post-translationally modified by sulfation, phosphorylation, and O-linked glycosylation, the sialylated sugar adduct being $HexNAcHex(NeuAc)_{(1-2)}$. Representative peptides containing one or more of the above-mentioned features are discussed in the following. For each of these representatives, we provide an overview of what is known from the literature as well as which novel instances of these classes we discovered in our CSF peptidome study.

1. Joining peptide (JP) of the corticotropin-lipotropin (pro-opiomelanocortin, POMC) precursor is a 30-amino-acid peptide located between the melanotropin γ and adrenocorticotropic hormone (ACTH) parts of the POMC precursor.[44,45] After cleavage at dibasic residues by proconvertases, C-terminal basic groups are removed by carboxypeptidase, and the glycine residue is acted upon by α-amidating monooxygenase, generating the C-terminally amidated peptide (Figure 2a). It was shown previously that JP is secreted as a homodimer through disulfide bonding between the single cysteine residues.[44] Illustrating how our technology can identify known neuropeptides in CSF, Figure 2b shows a representative MS/MS spectrum of the identified JP of POMC. The protein from which this peptide is derived, the corticotropin-lipotropin precursor, was not identified in the proteome profiling part of our work. This is as expected because it is being processed into different neuropeptides.

2. Neurexophilins are suggested to act as neuropeptides interacting with neurexins, a large family of neuronal cell surface proteins believed to be involved in intercellular signaling and formation of intercellular junctions.[46] We identified two novel peptide products derived from neurexophilin-3 and neurexophilin-4. Figure 3a shows the neurexophilin-3-originating peptide produced by removal of the signal peptide from the precursor and by proconvertase/carboxypeptidase action at the dibasic site. Mass spectrometry shows that the N-terminal glutamine of the peptide is converted into pyroglutamate. The conspicuous proline occurrence preceding and trailing basic residues probably not only influences peptide structure but also shields it from proteases.[47] Peptide sequence alignment shows a high level of conservation in mammals, including the signal peptide and dibasic cleavage sites, N-terminal glutamine, as well as the positioning of multiple proline residues (Figure 3b). Neurexophilins were not present in our CSF proteome map.

3. Synaptotagmins are believed to mediate calcium-dependent regulation of membrane trafficking. Only synaptotagmin-1 has been characterized in detail, and the functions of other members of the family are unclear.[48] We identified novel peptides derived from synaptotagmin-11, synaptotagmin-7, and synaptotagmin-4 as possible γ-secretase-cleavage products. Synaptotagmins were not identified in our CSF proteome profiling experiment. Figure 4a shows the synaptotagmin-11-derived peptide. The MS results reveal that the N-terminal

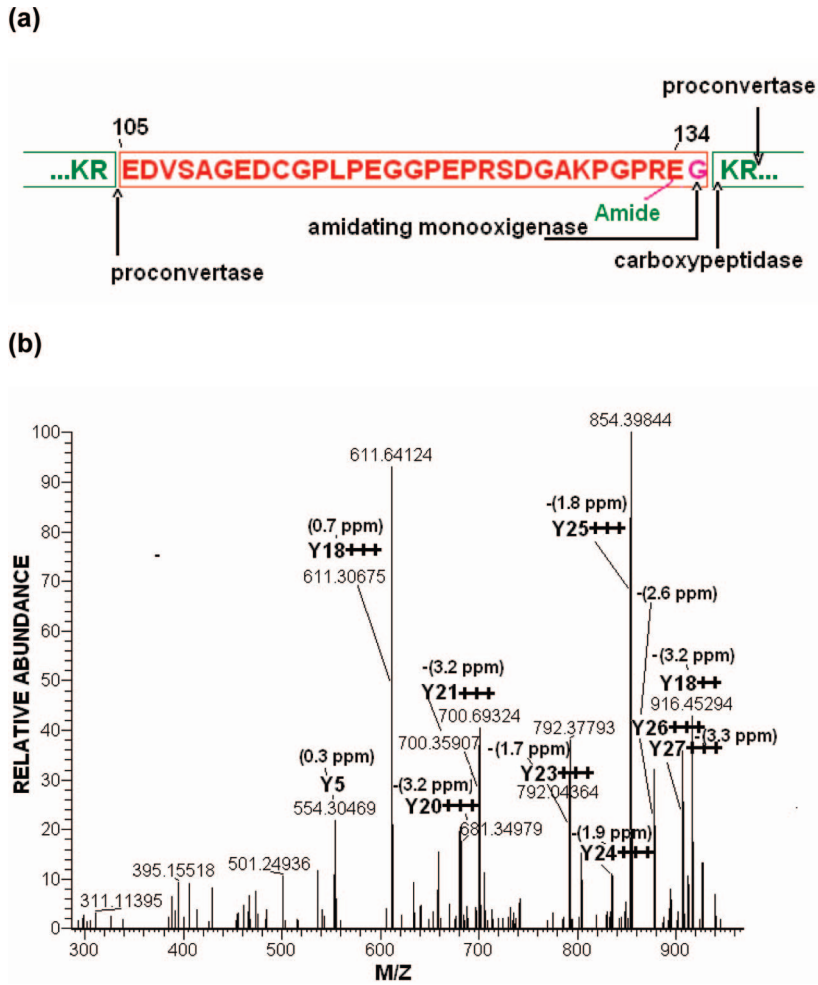


Figure 2. Joining peptide (JP) of pro-opiomelanocortin (POMC). (a) Suggested processing events leading to the formation of JP. (b) An example of the MS/MS fragmentation spectrum of the amidated *EDVSAGEDCGPLPEGGPEPRSDGAKPGPRE-NH2* peptide product (JP) of POMC. N-Terminal fragments (B-ions) and C-terminal fragments (Y-ions) are labeled. See ref 26 for an introduction to peptide sequencing. The charge states of multiply charged fragments are indicated by "plus" signs, and mass differences to calculated values are indicated in parentheses.

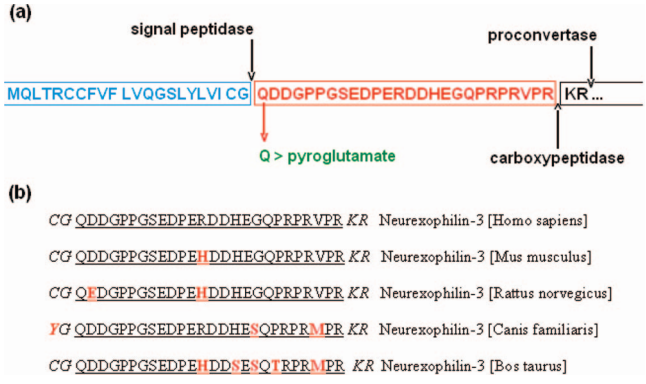


Figure 3. Neurexophilin-3-originating peptide. (a) Suggested processing events leading to the formation of the neurexophilin-3-derived peptide. (b) The alignment of the peptide sequence shows a high level of conservation of its sequence in mammals. Residues in red highlight differences from the human sequence.

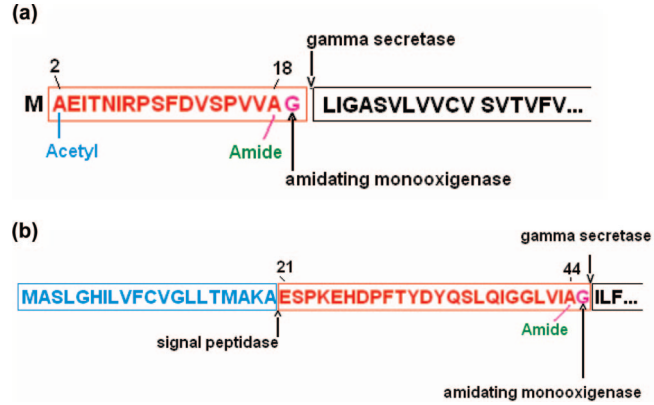


Figure 4. Possible processing events leading to the formation of the peptide derived from (a) synaptotagmin-11 and (b) phospholemman.

methionine of the precursor is removed and that alanine-2 is acetylated. Synaptotagmin-11 has a single-path transmembrane domain spanning residues 16–36. We find that γ-secretase cuts the precursor within the transmembrane domain at leucine-20, and glycine-19 is acted upon by α-amidating monooxygenase generating the C-terminally amidated peptide.

4. FXYD family members are single-span membrane proteins considered to either be regulators of ion channels or function as ion channels themselves.[49] We found peptides derived from FXYD-6 and FXYD-1 (phospholemman). Phospholemman is

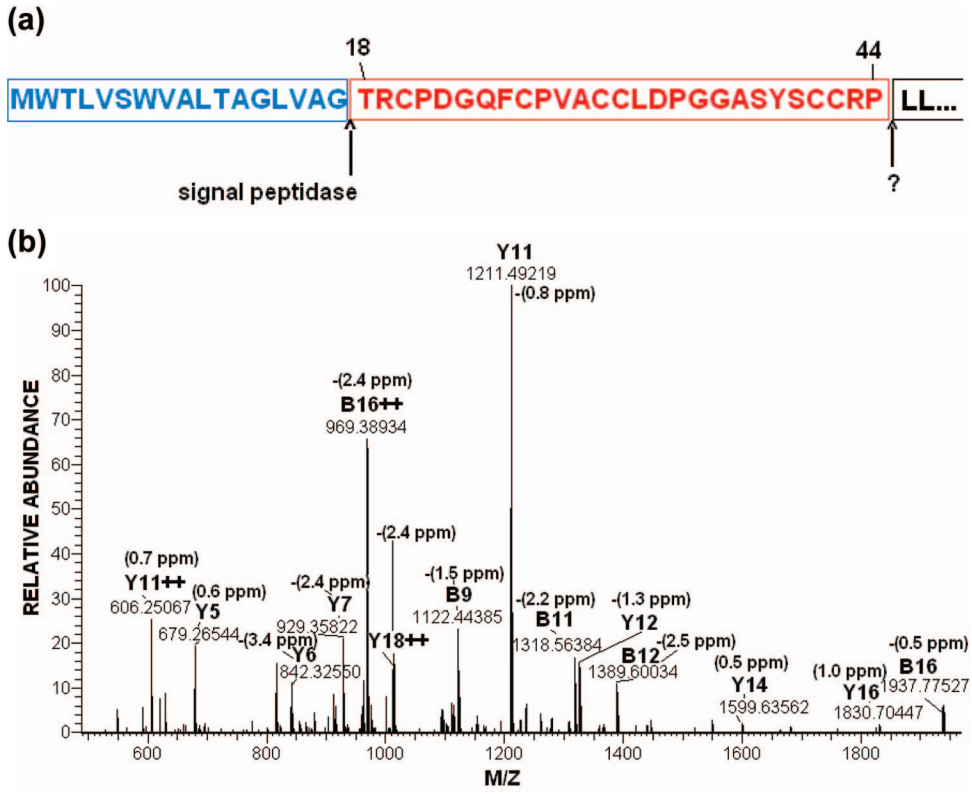


Figure 5. Paragranulin-like peptide. (a) Suggested processing events leading to the formation of the paragranulin-like peptide. (b) An example of the MS/MS fragmentation spectrum of the *RCPDGQFCPVACCLDPGGASYSCCRP* paragranulin-like peptide.

highly expressed in the CNS. It is most abundant in the cerebellum and Purkinje neurons and also enriched in choroid plexus.[50] Here we find that phospholemman is cleaved by γ-secretase or a similar enzyme. Figure 4b shows a phospholemman-derived peptide produced by removal of the signal sequence from the precursor and γ-secretase cleavage at isoleucine-46 within the transmembrane domain of phospholemman (residues 36−56). MS and MS/MS data show that the peptide is C-terminally amidated, indicating that glycine-45 is acted upon by α-amidating monooxygenase. FXYD proteins do not appear in our CSF proteome data set.

5. Granulins are cysteine-rich peptides which can modulate the growth of epithelial and mesenchymal cells in vitro. They are believed to be potential growth factors derived from the precursor glycoprotein and contain seven tandem repeats of the 12-cysteine granulin domain.[51] We identified one member of the family, a paragranulin-like peptide. This peptide, produced by action of signal peptidase at the N-terminus and an unknown enzyme at leucine-45 of the granulins precursor, spans residues 18−44. The C-terminal proline-44 directly preceding the basic arginine residues probably acts as a protection against protease cleavage. Paragranulin contains six cysteine residues likely used in intramolecular disulfide bond formation (Figure 5a). As an example of cysteine-containing peptides, Figure 5b presents an MS/MS spectrum of the paragranulin-like peptide. The granulin precursor protein was not identified in our CSF proteome profiling experiment.

6. Insulin-like growth factor-2 (IGF-2) is expressed in most embryonic tissues. After birth, IGF-2 expression ceases in some tissues, but the gene is often reactivated during tumorigenesis.[52] While the mature forms of insulin-like growth factors are structurally similar to insulin,[53] their functions and biogenesis are not completely understood. The production of IGF-2 involves removal of the signal peptide from the precursor followed by release of the mature 67-amino-acid peptide IGF-2 as well as the 89-amino-acid C-terminal E-peptide. The role of the E-peptide derived from pro-IGF-2 is unclear; however, the analogous E-peptide of pro-IGF1, another member of the IGF family, possesses mitogenic activity.[54] Here we identified, for the first time, glycosylated forms of peptides covering the first 21 residues of the E-peptide, as well as unmodified peptides originating from its middle part. IGF-2 peptides secreted by transfected HEK 293 cells contain significant amounts of sialic acid. The O-linked glycosylation sites of this overexpressed construct were mapped.[52]

O-Linked sugars are very labile and tend to "fall off" the peptide and "disappear" during the mass spectrometric fragmentation event. In the HCD fragmentation mode used here, an excess of the collision energy (CE, the potential difference with which the peptides are injected into the C-trap) leads to characteristic low mass ions (oxonium ions for sugars) as well as to peptide sequence specific ions, but information about the attachment site is lost. Figure 6 presents a fragmentation spectrum of a representative glycosylated peptide product of pro-IGF2. This peptide, with sequence *DVSTPPTVLPDNFPRYPVGKF*, is modified by O-linked glycosylation at the known site, threonine-99, carrying the $HexNAcHex(NeuAc)_2$ sugar. When the HCD fragmentation spectrum of the pro-IGF2-derived peptide was acquired with the CE value of 50, we observe a singly charged ion at m/z 274.092 corresponding to the dehydrated oxonium ion of sialic acid. We can also clearly follow the y-ion series, and taking into account the presence of the reporter ion indicating a possible sugar attachment, we identified the peptide (Figure 6). However, for the reasons

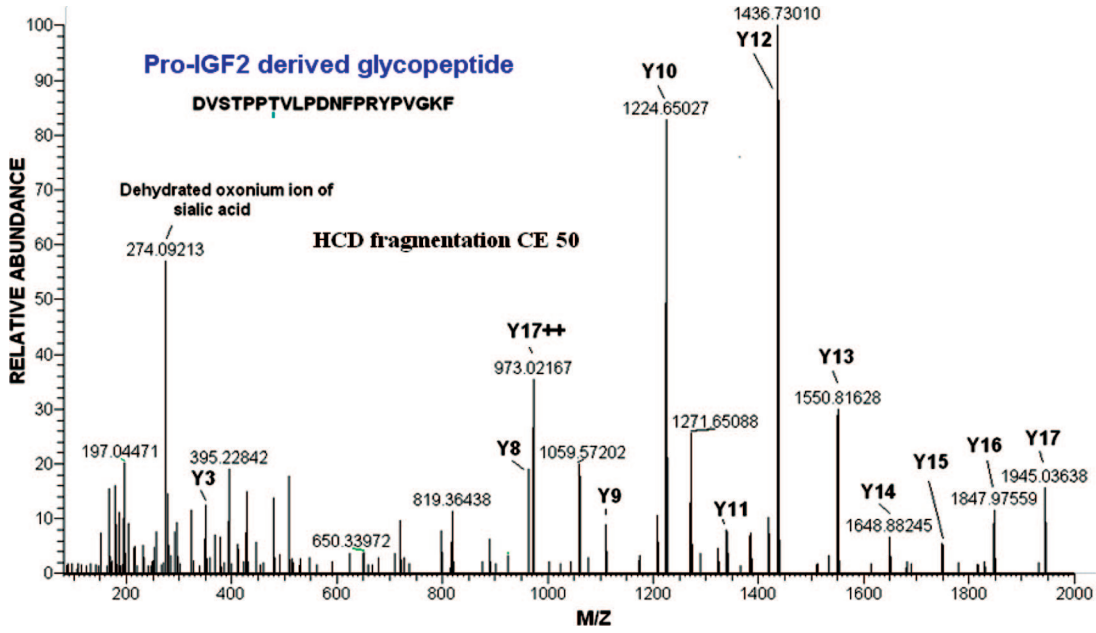


Figure 6. HCD fragmentation spectrum of the representative glycosylated *DVSTPPTVLPDNFPRYPVGKF* (O-linked $HexNAc_1Hex_1NeuAc_2$) peptide product of pro-IGF2 acquired with the CE value of 50.

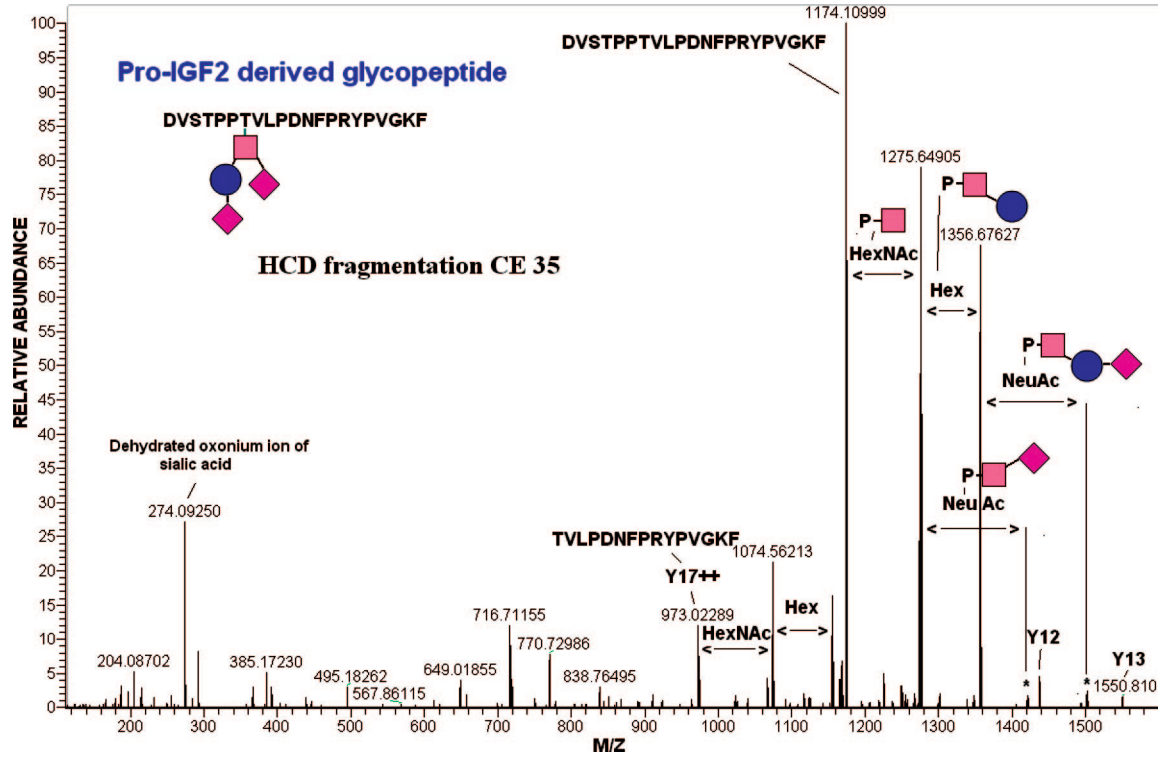


Figure 7. HCD fragmentation spectrum of the representative glycosylated *DVSTPPTVLPDNFPRYPVGKF* (O-linked $HexNAc_1Hex_1NeuAc_2$) peptide product of pro-IGF2 acquired with the CE value of 35.

noted above, the exact location of the modification, as well as its structure, is not apparent from this spectrum. We found that by lowering the CE from 50 to 35 we did observe fragment ions bearing sugar attachments (Figure 7). In this case, the sugars are partially preserved enabling us not only to pinpoint the O-linkage site but also to deduce the tentative glycosylation structure, even though the fragmentation of the peptide backbone does not yield many characteristic ions. Based on this identification of a glycopeptide, we modified the search engine to take the characteristic mass offsets due to glycosylation into account. The characteristic dehydration product of the sialic acid oxonium ion was then used to help verify the proposed identifications. In this way, we found fractalkine, SEL-OB, and Cadherin-20, solely as their glycosylated forms. We also identified the glycosylated peptide products of the heparin-binding EGF-like growth factor, which is discussed next.

7. Heparin-binding EGF-like growth factor (HB-EGF) precursor is a single-path transmembrane protein, containing a signal peptide, N-terminal pro-region, the actual HB-EGF

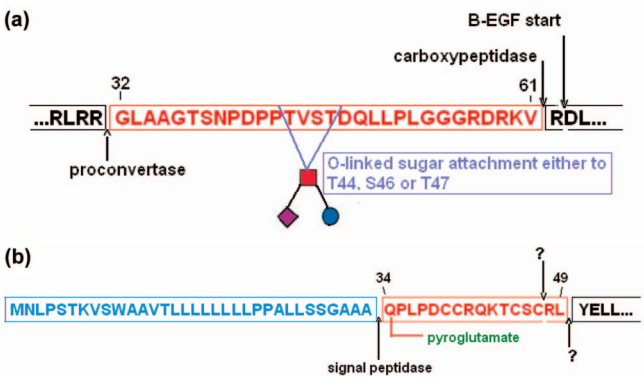


Figure 8. Suggested processing events leading to the formation of the peptides derived from (a) heparin-binding EGF-like growth factor and (b) orexin-A.

(comprised of heparin-binding and EGF-like domains, residues 63−148), juxtamembrane, transmembrane, and a short cytoplasmic part. HB-EGF binds to EGFR (ErbB1) and ErbB4 receptors and is mitogenic to fibroblasts and smooth muscle cells.[55,56] Even though the physiological role of HB-EGF is unclear, HB-EGF expression is essential for tumor formation in cancer-derived cell lines.[56] At least two O-linked glycosylation sites of HB-EGF are known.[55] We identified a novel glycosylated peptide derived from the N-terminal pro-region of the HB-EGF precursor. The peptide spans residues 32−61 and is produced by the action of carboxypeptidase and proconvertase enzymes (Figure 8a). We were not able to unequivocally pinpoint the site of glycosylation, but according to our MS/MS data (not shown), the $HexNAc_1Hex_1NeuAc_1$ sugar is attached to one of three residues: threonine-44, serine-46, or threonine-47.

8. Orexin-A and orexin-B (hypocretins) are a family of neuropeptide ligands for orexin GPCRs identified and named independently by two groups.[57,58] The peptide sequences are highly conserved in mammals. Their production takes place exclusively in the lateral hypothalamus, the area thought to be responsible for food intake. Nevertheless, the major known lesion associated with the malfunctioning of the orexin system is narcolepsy, a human sleep disorder caused by failure of orexin signaling due to impaired trafficking and processing of the precursor polypeptide.[59] Interestingly, while the orexin-A neuropeptide (residues 34–66 in the preprotein) has two disulfide bonds (Cys39−Cys45, Cys40−Cys47), there are no cysteine moieties in the orexin-B sequence (residues 70–97 in the preprotein). This implies different 3D structures and interaction profiles. Indeed, the orexin-1 receptor has significantly stronger affinity for orexin-A, whereas the orexin-2 receptor binds both neuropeptides with similar affinity.[59] In this work, we identified two cysteine-containing peptide variants (residues 34–47 and 34–49) derived from the N-terminal part of orexin-A (Figure 8b). Both peptides are produced by signal peptidase action on the N-terminal part of the precursor protein. MS shows that the N-terminal glutamine of the peptides is transformed into pyroglutamic acid. The origin of enzymes participating in the C-terminal cleavages is unclear. The quadruple cysteine structure is preserved in both peptides.

Apart from the examples given above, our CSF peptidome data contain a wealth of other information on potential regulatory neuropeptides as well as unexpected findings such as a number of *polyproline* peptides of unknown origin containing 7−12 proline residues. To facilitate use of the data by the community, we have made the CSF peptidome available as part of the MAPU database.

CSF Proteome. As mentioned above, our CSF proteome study utilized samples from six individuals, of which one was profiled independently and the other five were profiled as a pooled sample. For the profiling of the individual sample, we used methods similar to those in our previous seminal, tear fluid, and urinary proteome studies.[60−62] In particular, two consecutive stages of fragmentation were employed on a linear ion trap Fourier transform instrument (LTQ-FT) to essentially eliminate all false positive identifications.[63] The pooled sample was analyzed with the LTQ-Orbitrap, again requiring extremely stringent identification criteria with at least two peptides required for positive protein validation. A reverse database search indicates that using the employed criteria our CSF proteome should not contain any false positive protein identifications (see Methods). To combine the proteome profiling experiments and to analyze the data, we used the ProteinCenter software (www.proxeon.com). The overall features of our CSF proteome map (Supporting Information Table 5) were furthermore classified using Gene Ontology (GO)[64,65] for protein categorization (see Methods and Supporting Information Figure 1).

We first compared the proteome features between single (657 identified proteins) and pooled (531 identified proteins) samples to make sure that individual sample variability had no influence on the general proteome profile. As shown in Figure 9, the proportions of characteristic features such as the number of membrane and extracellular proteins, proteins engaged in signal transduction, and receptor activity as well as proteins with a predicted signal sequence are very similar between the two data sets. From this, we conclude that our sample of six individuals is representative of the CSF proteome. GO analysis of our CSF proteome shows that by far the major predicted cellular localizations are membrane and extracellular (44 and 38%, respectively; note that proteins can have several GO localizations). There are few proteins with annotated intracellular localizations; for example, Golgi, cytosol, and nuclear categories account for less than 4% each. Additionally, the proportion of proteins predicted by the SignalP algorithm to carry signal peptides was 74%, whereas it is only 15% for the proteome as a whole. While largely expected, these findings support the high quality of the CSF preparation and analysis. GO functional analysis revealed a striking proportion of proteins involved in signal transducer and receptor activities (25 and 17%, respectively). We next determined the overlap between our CSF proteome map and the recently published CSF proteome data set of Abdi et al.[8] From that data set, we extracted proteins identified with at least two peptides. With the aid of ProteinCenter, proteins in both data sets were clustered at the 98% homology level to group alleles of the same proteins with fragments and to remove redundancy in protein sequences. Additionally, keratin identifications were filtered out. The overlap between our data set (767 proteins) and the CSF data set of Abdi et al. (650 proteins) was close to 40%. This overlap is reasonable, in our opinion, as different methodologies were employed and as neither CSF proteome is exhaustive at this stage. We further compared our proteome to the recently published CSF data set of Pan et al.[66] That CSF proteome was created by the combination of a large number of CSF studies

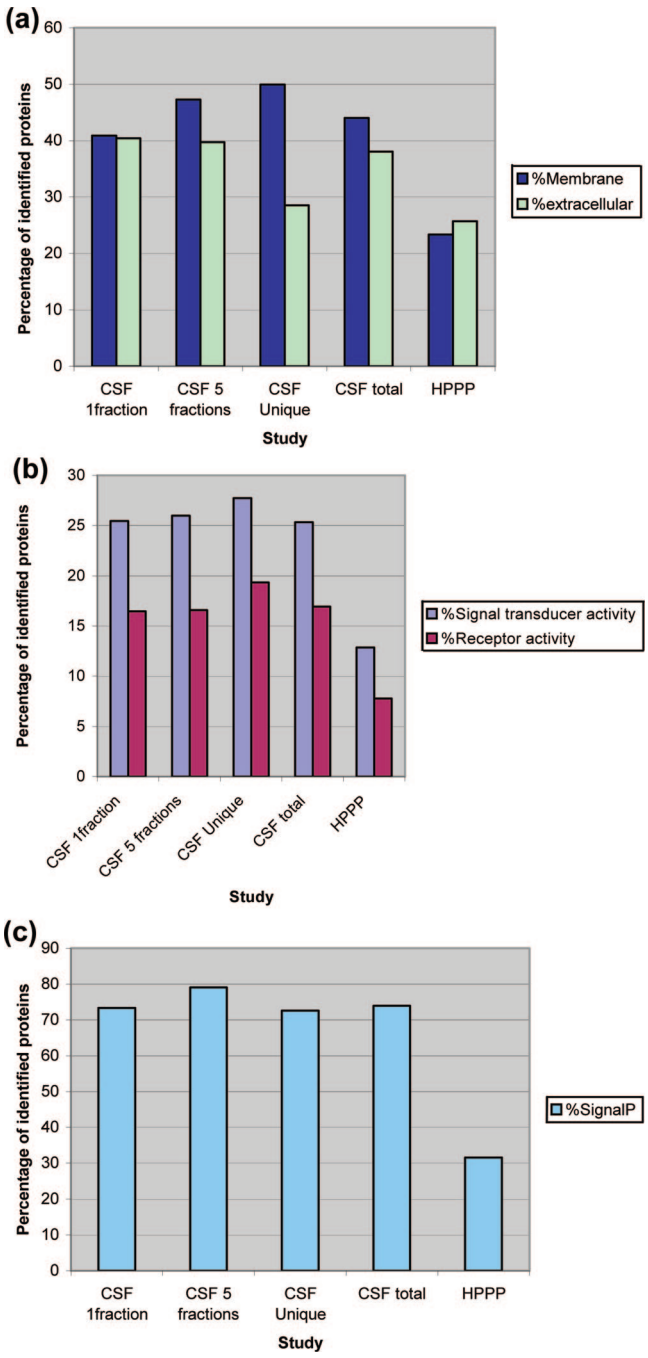


Figure 9. Comparison of our CSF proteome with the HUPO plasma data set published by States et al. (a) GO annotation of membrane and extracellular localizations for the in-depth CSF proteome of one individual, the pooled CSF proteome of five individuals, the proteome obtained by subtraction of the HUPO plasma data set from the total CSF data set, the total CSF data set, and the HUPO plasma data set. (b) GO functional annotation for signal transducer and receptor activity. (c) SignalP prediction of proteins with signal peptides.

using different methodologies. We found 75% overlap of our data with theirs at the 98% homology level.

Comparison of the CSF Proteome with the Human Plasma Proteome. Cerebrospinal fluid and plasma are specialized body fluids bridged at the choroid plexus. Plasma proteins enter CSF by a simple diffusion-based mechanism as a function of their molecular size across the blood−brain barrier. It is generally believed that the majority of proteins in CSF originate from blood.[67] We compared our data with the published Human Proteome Organization (HUPO) high-confidence plasma proteome data set of 889 proteins[68] using similar methodology as described above. We found that 581 proteins were unique to our CSF data set (76% of the CSF proteome). When comparing the identified proteins in each study for their GO cellular compartment annotation, we found that extracellular and membrane-bound proteins were much more abundant in the CSF proteome than in the HUPO plasma proteome (Figure 9a). We subtracted the HUPO plasma proteome from our data set, creating a CSF "unique" proteome. Note that the plasma proteome is currently incomplete and that further quantitative studies would be necessary to define truly unique CSF proteins. Interestingly, the portion of membrane-bound and extracellular proteins was similar between plasma and "CSF unique" proteomes (23% and 26%, respectively), whereas the CSF unique proteome has more than twice the proportion of membrane-bound proteins (Figure 9a). Remarkably, about 74% of all CSF proteins have a predicted signal peptide sequence, while only 32% of the HUPO plasma proteins do (Figure 9c). Similarly, the proportion of proteins annotated for signal transduction and for receptor activity was much higher in CSF than in plasma (26% vs 13% and 17% vs 8%, respectively; see Figure 9b). Supporting the findings from our CSF proteome study, we found that the overlap of the Abdi et al. CSF data set with the HUPO plasma proteome was similar (21% vs 24% in our data set; Supporting Information Figure 3). Based on this observation, we conclude that even though a significant share of proteins in CSF appears to originate from plasma, the intrinsic CSF proteins constitute the major part of its proteome, at least at the level detected by current proteome profiling technologies.

Origin of Membrane Proteins in CSF. The most outstanding feature of the CSF proteome is the prevalence of membrane-bound proteins. The CSF peptidome analysis described above provides at least a partial explanation of this finding. It is well-known that proteolysis is responsible not only for remodeling of the cell surface but also for release of some of the membrane-bound proteins from their anchor.[69,70] Based on our observation of the cleavage sites of neuropeptides and the extracellular parts of the corresponding proteins (see below), we suggest that a number of the identified membrane-bound proteins are proteolytically cleaved in the membrane vicinity and their extracellular parts are released into the surroundings. The situation appears to be different from that in the urinary proteome where we observed peptides from the full lengths of the proteins[60] and where secretion in exosomes has been proposed to be the major process responsible for the presence of membrane proteins in this body fluid. Below we discuss several specific examples of the proposed proteolytic mechanism.

1. Amyloid precursor-like protein 1 (APLP1), β-amyloid precursor protein (APP), and amyloid precursor-like protein 2 (APLP2) belong to a family of type I membrane proteins. All three proteins are closely related and are subject to site-specific proteolysis by secretases.[71,72] Even though the exact function of APLP1 is unknown, recent studies have shown that it modulates endocytosis and proteolytic processing of APP.[73] A possible role of APLP1 in regulation of α2A-adrenergic receptor trafficking has also been proposed.[74] The sequence of APLP1 is comprised of a signal peptide (residues 1–38), an extracellular part (residues 39–650), a membrane part (residues 581–603), and a short C-terminal cytoplasmic part (residues 604–650). We identified the extracellular part of APLP1 during our CSF

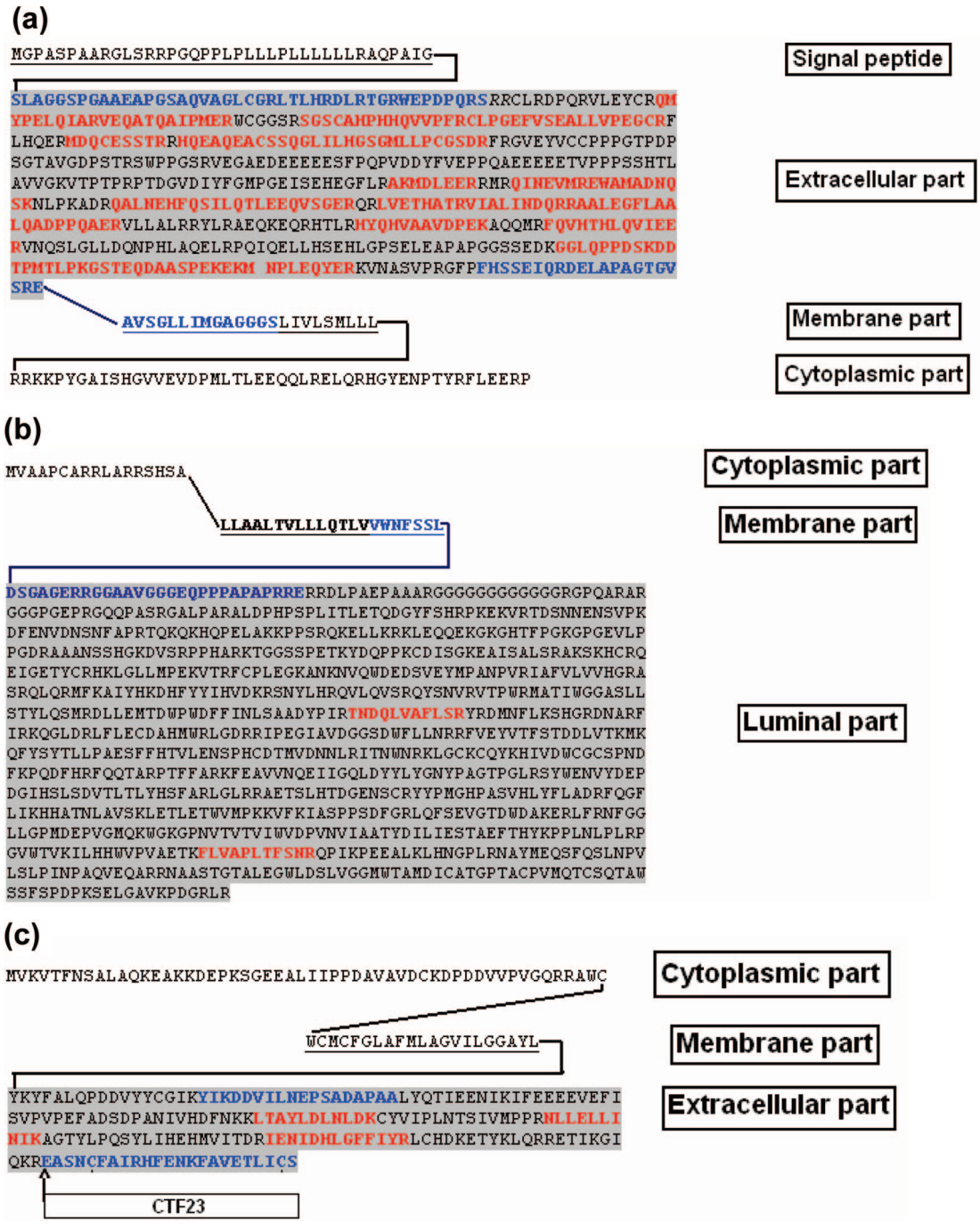


Figure 10. Interplay of proteome and peptidome data. (a) Sequence coverage of amyloid precursor-like protein 1; (b) xylosyltransferase I; (c) integral membrane protein 2B. Sequence areas identified during the CSF proteome mapping experiment are shown in red. Sequence areas identified during the CSF peptidome profiling are shown in blue.

proteome analysis. During CSF peptidome profiling, we found peptides confirming the secretase-related intramembrane cleavage of APLP1 and a peptide (residues 39–81) originated from the removal of the signal peptide (residues 1–38) and convertase action at the dibasic site (residues 82–83). The identified protein sequence covers the region between the convertase cleavage site (82–83) and the near-membrane region of the protein (Figure 10a). The near-membrane secretase cleavage is the probable cause of APLP1's launch into extracellular space.

2. Xylosyltransferase I (XT-I) catalyzes transfer of xylose from UDP-xylose to serine residues in proteoglycan core proteins and is a regulatory factor in chondroitin sulfate synthesis.[75] XT-I is a type II membrane protein with a short N-terminal part (residues 1–17) facing the cytoplasm, signal anchor (residues 18–38), and C-terminal part (residues 39–959) in the luminal part of the rough endoplasmic reticulum. Secretion of XT-I into the extracellular space has already been shown.[76] XT-I was proposed as a synovial fluid marker of cartilage destruction[77] and, more recently, as a serum marker in systemic sclerosis.[78] We identified XT-I during the CSF proteome mapping experiment. Additionally, peptides originating from the signal-anchor membrane part of XT-I were also identified in the CSF peptidome (Figure 10b). These peptides carry the intramembrane secretase-like cleavage signature at their N-termini and

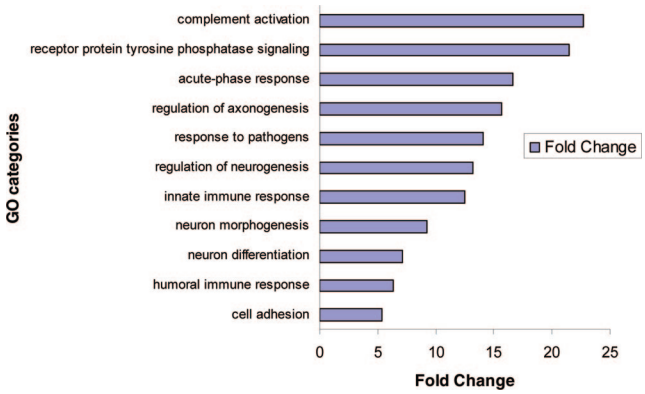


Figure 11. Statistically over-represented GO terms in our CSF proteome data set as compared to annotations for the entire human proteome.

dibasic cleavage signature at their C-termini. It is likely that with the signal anchor cut the C-terminal part of the protein is released into extracellular space.

3. Integral membrane protein 2B (ITM2B also known as BRI2 or E25B) is a type II integral membrane protein that belongs to a family including two other members, ITM2A (BRI1 or E25A) and ITM2C (BRI3 or E25C). The extracellular part of BRI2 is known to be proteolytically processed by convertases resulting in the release of a 23-amino-acid C-terminal fragment of unknown function (CTF23).[79] Mutations in the ITM2B gene are associated with familial British and Danish forms of dementia.[80,81] ITM2B interacts with amyloid precursor protein (APP) and has a modulatory effect on APP processing, increasing the levels of cellular APP as well as its γ-secretase-generated C-terminal fragments.[82] The extracellular part of ITM2B was identified during the CSF proteome profiling. CTF-23 and CTF-23-derived peptides were identified during CSF peptidome mapping (Figure 10c). Additionally, the 18-amino-acid peptide *YIKDDVILNEPSADAPAA* derived from the membrane proximal region was also identified. It is possible that the production of this peptide leads to the release of the extracellular part of ITM2B.

Overrepresentation of Transmembrane Receptor Protein Tyrosine Phosphatases. We used the BinGO tool[27] to find statistically over-represented GO terms in our CSF proteome data set in comparison to the GO annotations of the whole human proteome (Figure 11). In addition to the significant enrichment of basic CSF functions related to host defense and intrinsic neurological regulation, the transmembrane receptor protein tyrosine phosphatase (PTPR) term was also significantly augmented ($p = 1.95 \times 10^{-7}$). The functional meaning of this PTPR enrichment is unknown; however, proteome and peptidome profiling data shed some light on the mechanisms leading to the presence of these proteins in CSF as will be discussed below.

Protein tyrosine phosphatase receptor type Z (PTPRZ) is a member of the PTPR subgroup of the protein phosphatase family. Even though tyrosine phosphatases are known to be signaling molecules that regulate a variety of developmental and functional processes in the CNS, the intracellular mechanisms by which they regulate cellular signaling pathways are not well understood.[83] PTPRZ is specifically expressed in the central nervous system. It is a single-pass type I membrane protein with two cytoplasmic tyrosine-protein phosphatase domains and an α-carbonic anhydrase and a fibronectin type III extracellular domain. The protein sequence includes a 24-amino-acid signal peptide at the N-terminus, a large extracellular part (residues 25–1635), a 26-amino-acid membrane part (residues 1636–1661), and a 653-amino-acid C-terminal cytoplasmic part (residues 1662–2314). It has been reported that PTPRZ plays a critical role in functional recovery from demyelinating lesions in multiple sclerosis.[84] Additionally, it has been suggested that PTPRZ may regulate glioblastoma cell motility.[85] We identified the extracellular part of PTPRZ during a CSF proteome mapping experiment and peptides originating from the membrane part of PTPRZ during CSF peptidome mapping (Supporting Information Figure 4). These peptides carry the intramembrane secretase-like cleavage signature at their C-termini. It is likely that with the membrane part of PTPRZ cut its extracellular part is released into extracellular space. Taking into account the specific expression of PTPRZ in remyelinating oligodendrocytes in multiple sclerosis lesions, it is tempting to suggest that the soluble form of PTPRZ and/or the peptides originated from the PTPRZ secretase-like cleavage could be used as potential markers in multiple sclerosis.

The extracellular parts of other PTPRs were also identified in our CSF proteome mapping experiment. These include Islet antigen-2 (IA-2 or PTPR-like N protein), an enzymatically inactive PTPR located in neuroendocrine cells throughout the body, which is a major autoantigen in type 1 diabetes[86] (Supporting Information Figure 2a). Receptor tyrosine phosphatase δ, another example, is thought to be a neurite-promoting homophilic adhesion molecule[87] (Supporting Information Figure 2b). Even though the corresponding peptides carrying the intramembrane cleavage signature were identified only for PTPRZ, we suggest that the mechanisms responsible for release of the extracellular parts of the other PTPR group members are similar to that proposed for PTPRZ.

Bioinformatics analysis of the CSF proteome furthermore identified a large number of protein peptidases. In total, 42 hydrolases (EC 3.4) including 6 carboxypeptidases, 3 convertases, and 5 aminopeptidases were detected. The occurrence of these enzymes in the CSF samples hints at their potential involvement in post-translational truncations of CSF proteins. However, as is the case for all the enzymes discussed here, it remains to be determined whether they and their associated peptides are mere byproducts of the process of remodeling and rearranging the cell surface or whether they play specific functional roles in the mostly unexplored CSF universe.

Conclusions

We believe that the presented data set contains valuable unique information which will enable interested researchers to identify novel CNS regulatory mechanisms. Our study demonstrates that MS-based technology has already advanced sufficiently to allow not only detection of large numbers of proteins of different abundance in biological fluids but also identification of their endogenously processed peptides. Thus, MS can now be used to study these processing events resulting in cleavage or changing of the cleavage pattern which may allow linking them to specific physiological or pathophysiological events. Our data suggest that the major contributors to the protein population of CSF are protein secretion combined with ongoing proteolytic processes involved in cell surface remodeling, protein shedding, and creation of regulatory peptides. The CSF protein population differs in composition from that of

plasma presumably due to inherent CSF functions of an as yet mostly unexplored nature.

The identification of 563 peptide forms and 798 proteins in our CSF profiling study offers a library to the community working on cerebrospinal fluid and endocrine signaling. To enable unhindered and open access to our data, we made them available at the Max-Planck Unified Proteome database (MAPU).[88] With time and further research, we expect the cerebrospinal fluid to be an important source of novel regulatory peptides and biological markers.

Abbreviations: ABC, ammonium bicarbonate; ACTH, adrenocorticotropic hormone; CNS, central nervous system; APLP1, amyloid precursor-like protein 1; APLP2, amyloid precursor-like protein 2; APP, β-amyloid precursor protein; BiNGO, Biological Networks Gene Ontology; CE, collision energy; CSF, cerebrospinal fluid; FT-ICR, Fourier transform ion cyclotron resonance; FTMS, Fourier transform mass spectrometry; GeLC−MS, gel enhanced liquid chromatography–mass spectrometry; GO, Gene Ontology; GPCR, G protein-coupled receptor; HB-EGF, Heparin-binding EGF-like growth factor; HCD, higher-energy C-trap dissociation; HPPP, Human Plasma Proteome Project; HUPO, Human Proteome Organization; IGF-2, insulin-like growth factor-2; ITM2B, integral membrane protein 2B; JP, joining peptide of POMC; LC-MS/MS, liquid chromatography-tandem mass spectrometry; LTQ-FT, linear quadrupole ion trap-Fourier transform mass spectrometer; MS, mass spectrometry; MS/MS, tandem mass spectrometry; MS^3, MS/MS/MS; POMC, pro-opiomelanocortin; PTPR, protein tyrosine phosphatase receptor; PTPRZ, protein tyrosine phosphatase receptor type ζ; SDS-PAGE, sodium dodecyl sulfate-polyacrylamide gel electrophoresis; SIM, selected ion monitoring; STAGE, stop and go extraction; TFA, trifluoroacetic acid; XT-I, xylosyltransferase I.

Supporting Information Available: CSF peptides identified by high-accuracy mass spectrometry in both MS and MS/MS modes (Supplementary Table 1). Unique peptide precursors identified solely during CSF peptidome profiling (Supplementary Table 2). Identified CSF peptides derived by convertase cleavage (Supplementary Table 3). Identified CSF peptides derived by secretase cleavage (Supplementary Table 4). CSF protiens identified by high-accuracy mass spectrometry (Supplementary Table 5). CSF proteome GO categorization (Supplementary Figure 1). Sequence coverage of *PTPR-like N* and *PTPR-delta* at the proteome level (Supplementary Figure 2). Comparision of CSF proteome with Human Plasma Proteome (Supplementary Figure 3). Sequence coverage of receptor tyrosine phosphatase zeta at the proteome and peptidome level (Supplementary Figure 4). This material is available free of charge via the Internet at http://pubs.acs.org.

Acknowledgment. The authors thank Dr. Yong Zhang for technical advice. Work at the Center of Experimental Bioinformatics (CEBI) is supported by a generous grant from the Danish National Research Foundation. We are also grateful to our colleagues at the Department of Proteomics and Signal Transduction, Max-Planck-Institute, for helpful discussions.

References

(1) Brumback, R. A. Anatomic and physiologic aspects of the cerebrospinal fluid space. In *The Cerebrospinal Fluid*; Herndon, R. M., Brumback, R. A., Eds.; Kluwer: Boston, 1989; pp 15–43.

(2) Thompson, E. J.; Keir, G. Laboratory investigation of cerebrospinal fluid proteins. *Ann. Clin. Biochem.* **1990**, *27* (Pt 5), 425–35.

(3) Yuan, X.; Desiderio, D. M. Proteomics analysis of human cerebrospinal fluid. *J. Chromatogr. B, Analyt Technol. Biomed. Life Sci.* **2005**, *815* (1–2), 179–89.

(4) Aebersold, R.; Goodlett, D. R. Mass spectrometry in proteomics. *Chem. Rev.* **2001**, *101* (2), 269–95.

(5) Geho, D. H.; Liotta, L. A.; Petricoin, E. F.; Zhao, W.; Araujo, R. P. The amplified peptidome: the new treasure chest of candidate biomarkers. *Curr. Opin. Chem. Biol.* **2006**, *10* (1), 50–5.

(6) Puchades, M.; Hansson, S. F.; Nilsson, C. L.; Andreasen, N.; Blennow, K.; Davidsson, P. Proteomic studies of potential cerebrospinal fluid protein markers for Alzheimer's disease. *Brain Res. Mol. Brain Res.* **2003**, *118* (1–2), 140–6.

(7) Zhang, J.; Goodlett, D. R.; Quinn, J. F.; Peskind, E.; Kaye, J. A.; Zhou, Y.; Pan, C.; Yi, E.; Eng, J.; Wang, Q.; Aebersold, R. H.; Montine, T. J. Quantitative proteomics of cerebrospinal fluid from patients with Alzheimer disease. *J. Alzheimers Dis.* **2005**, *7* (2), 125–33, discussion 173–80.

(8) Abdi, F.; Quinn, J. F.; Jankovic, J.; McIntosh, M.; Leverenz, J. B.; Peskind, E.; Nixon, R.; Nutt, J.; Chung, K.; Zabetian, C.; Samii, A.; Lin, M.; Hattan, S.; Pan, C.; Wang, Y.; Jin, J.; Zhu, D.; Li, G. J.; Liu, Y.; Waichunas, D.; Montine, T. J.; Zhang, J. Detection of biomarkers with a multiplex quantitative proteomic platform in cerebrospinal fluid of patients with neurodegenerative disorders. *J. Alzheimers Dis.* **2006**, *9* (3), 293–348.

(9) Dumont, D.; Noben, J. P.; Raus, J.; Stinissen, P.; Robben, J. Proteomic analysis of cerebrospinal fluid from multiple sclerosis patients. *Proteomics* **2004**, *4* (7), 2117–24.

(10) Conti, A.; Sanchez-Ruiz, Y.; Bachi, A.; Beretta, L.; Grandi, E.; Beltramo, M.; Alessio, M. Proteome study of human cerebrospinal fluid following traumatic brain injury indicates fibrin(ogen) degradation products as trauma-associated markers. *J. Neurotrauma* **2004**, *21* (7), 854–63.

(11) Siman, R.; McIntosh, T. K.; Soltesz, K. M.; Chen, Z.; Neumar, R. W.; Roberts, V. L. Proteins released from degenerating neurons are surrogate markers for acute brain damage. *Neurobiol. Dis.* **2004**, *16* (2), 311–20.

(12) Zhang, J.; Goodlett, D. R.; Peskind, E. R.; Quinn, J. F.; Zhou, Y.; Wang, Q.; Pan, C.; Yi, E.; Eng, J.; Aebersold, R. H.; Montine, T. J. Quantitative proteomic analysis of age-related changes in human cerebrospinal fluid. *Neurobiol. Aging* **2005**, *26* (2), 207–27.

(13) Yuan, X.; Russell, T.; Wood, G.; Desiderio, D. M. Analysis of the human lumbar cerebrospinal fluid proteome. *Electrophoresis* **2002**, *23* (7–8), 1185–96.

(14) Yuan, X.; Desiderio, D. M. Proteomics analysis of prefractionated human lumbar cerebrospinal fluid. *Proteomics* **2005**, *5* (2), 541–50.

(15) Finehout, E. J.; Franck, Z.; Lee, K. H. Towards two-dimensional electrophoresis mapping of the cerebrospinal fluid proteome from a single individual. *Electrophoresis* **2004**, *25* (15), 2564–75.

(16) Davidsson, P.; Paulson, L.; Hesse, C.; Blennow, K.; Nilsson, C. L. Proteome studies of human cerebrospinal fluid and brain tissue using a preparative two-dimensional electrophoresis approach prior to mass spectrometry. *Proteomics* **2001**, *1* (3), 444–52.

(17) Stark, M.; Danielsson, O.; Griffiths, W. J.; Jornvall, H.; Johansson, J. Peptide repertoire of human cerebrospinal fluid: novel proteolytic fragments of neuroendocrine proteins. *J. Chromatogr. B, Biomed. Sci. Appl.* **2001**, *754* (2), 357–67.

(18) Yuan, X.; Desiderio, D. M. Human cerebrospinal fluid peptidomics. *J. Mass Spectrom.* **2005**, *40* (2), 176–81.

(19) Reiber, H.; Peter, J. B. Cerebrospinal fluid analysis: disease-related data patterns and evaluation programs. *J. Neurol. Sci.* **2001,** *184* (2), 101–22.

(20) Thomas, L. *Labor und Diagnose Indikation und Bewertung von Laborbefunden für die medizinische Diagnostik*, 6th ed.; TH-Books Verlagsgesellschaft:frankfurt/main, 2005.

(21) Kluge, H.; Wieczorek, V.; Linke, E.; Zimmermann, K.; Isenmann, S.; Witte, W. O. *Atlas of CSF Cytology*, 1st ed.; Georg Thieme Verlag: Stuttgart/New York, 2007.

(22) Rappsilber, J.; Ishihama, Y.; Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **2003**, *75* (3), 663–70.

(23) Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-trap. *Mol. Cell. Proteomics* **2005**, *4* (12), 2010–21.

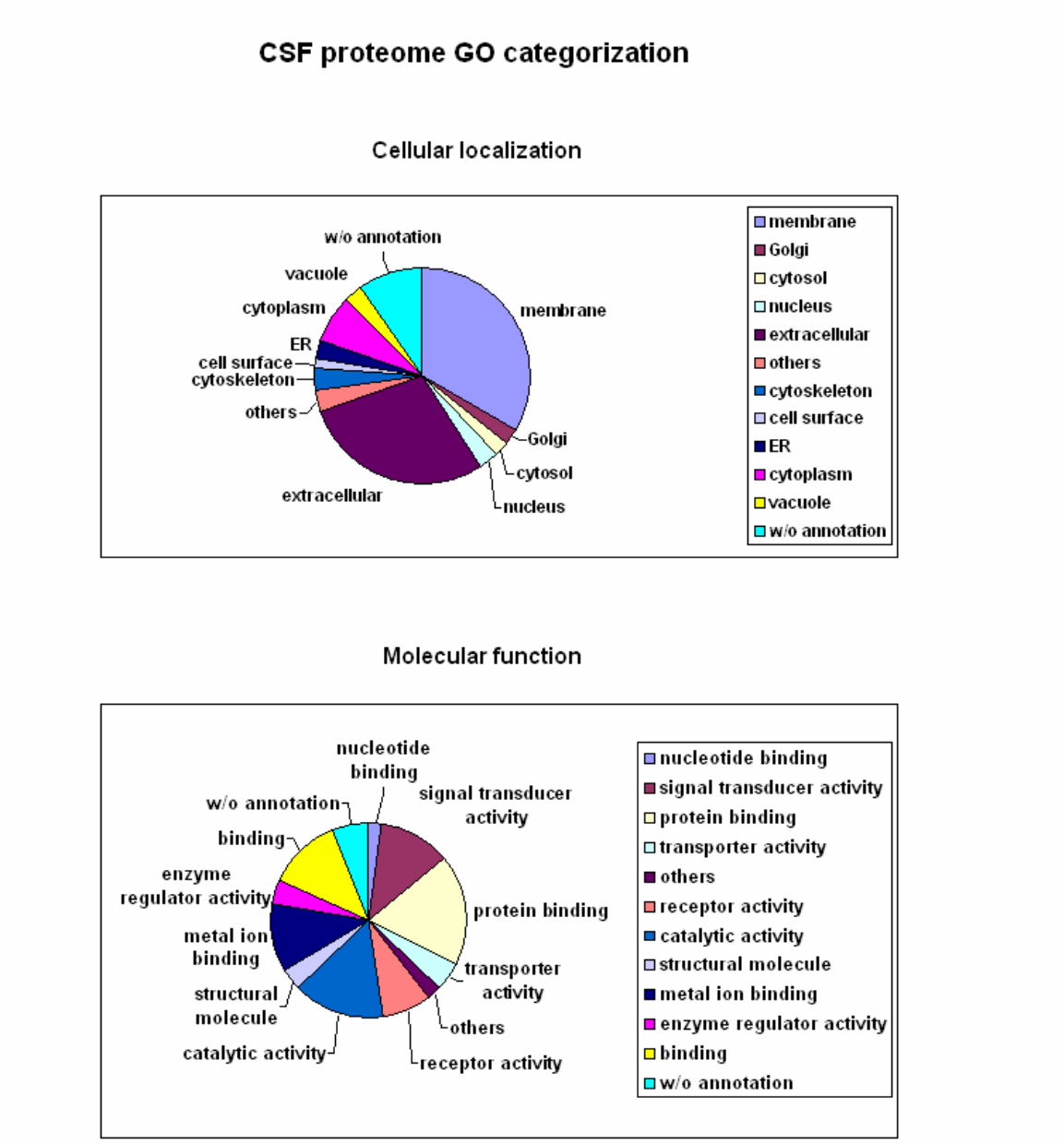
(24) Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **2004**, *4* (7), 1985–8.

(25) 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* **1999**, *20* (18), 3551–67.
(26) Steen, H.; Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **2004**, *5* (9), 699–711.
(27) Maere, S.; Heymans, K.; Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **2005**, *21* (16), 3448–9.
(28) European Bioinformatics Institute Gene Ontology Annotation (GOA) Database. http://www.ebi.ac.uk/GOA/.
(29) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **1995**, (B57), 289–300.
(30) de Godoy, L. M.; Olsen, J. V.; de Souza, G. A.; Li, G.; Mortensen, P.; Mann, M., Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.* **2006**, *7* (6), R50.
(31) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **2007**, *4* (9), 709–12.
(32) Southey, B. R.; Amare, A.; Zimmerman, T. A.; Rodriguez-Zas, S. L.; Sweedler, J. V. NeuroPred: a tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides. *Nucleic Acids Res.* **2006**, *34* (Web Server issue), W267–72.
(33) Bergeron, F.; Leduc, R.; Day, R. Subtilase-like pro-protein convertases: from molecular specificity to therapeutic applications. *J. Mol. Endocrinol.* **2000**, *24* (1), 1–22.
(34) Eipper, B. A.; Milgram, S. L.; Husten, E. J.; Yun, H. Y.; Mains, R. E. Peptidylglycine alpha-amidating monooxygenase: a multifunctional protein with catalytic, processing, and routing domains. *Protein Sci.* **1993**, *2* (4), 489–97.
(35) Desouza, E. B.; Kuhar, M. J. Corticotropin-releasing factors receptors in the pituitary gland and central nervous system: Methods and overview. *Methods Enzymol.* **1986**, *124*, 560–590.
(36) Fricker, L. D. Neuropeptide-processing enzymes: applications for drug discovery. *Aaps J.* **2005**, *7* (2), E449–55.
(37) Abraham, G. N.; Podell, D. N. Pyroglutamic acid: nonmetabolic formation, function in proteins and peptides, and characteristics of the enzymes effecting its removal. *Mol. Cell. Biochem.* **1981**, *38*, 181–190.
(38) Hinkle, P. M.; Tashjian, A. H., Jr. Receptors for thyrotropin-releasing hormone in prolactin producing rat pituitary cells in culture. *J. Biol. Chem.* **1973**, *248* (17), 6180–6.
(39) Neitz, S.; Jurgens, M.; Kellmann, M.; Schulz-Knappe, P.; Schrader, M. Screening for disulfide-rich peptides in biological sources by carboxyamidomethylation in combination with differential matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2001**, *15* (17), 1586–92.
(40) Vitt, U. A.; Hsu, S. Y.; Hsueh, A. J. Evolution and classification of cystine knot-containing hormones and related extracellular signaling molecules. *Mol. Endocrinol.* **2001**, *15* (5), 681–94.
(41) Mygind, P. H.; Fischer, R. L.; Schnorr, K. M.; Hansen, M. T.; Sonksen, C. P.; Ludvigsen, S.; Raventos, D.; Buskov, S.; Christensen, B.; De Maria, L.; Taboureau, O.; Yaver, D.; Elvig-Jorgensen, S. G.; Sorensen, M. V.; Christensen, B. E.; Kjaerulff, S.; Frimodt-Moller, N.; Lehrer, R. I.; Zasloff, M.; Kristensen, H. H. Plectasin is a peptide antibiotic with therapeutic potential from a saprophytic fungus. *Nature* **2005**, *437* (7061), 975–80.
(42) Nyborg, A. C.; Ladd, T. B.; Zwizinski, C. W.; Lah, J. J.; Golde, T. E. Sortilin, SorCS1b, and SorLA Vps10p sorting receptors, are novel gamma-secretase substrates. *Mol. Neurodegener.* **2006**, *1*, 3.
(43) Kang, J.; Lemaire, H. G.; Unterbeck, A.; Salbaum, J. M.; Masters, C. L.; Grzeschik, K. H.; Multhaup, G.; Beyreuther, K.; Muller-Hill, B. The precursor of Alzheimer's disease amyloid A4 protein resembles a cell-surface receptor. *Nature* **1987**, *325* (6106), 733–6.
(44) Bertagna, X.; Camus, F.; Lenne, F.; Girard, F.; Luton, J. P. Human joining peptide: a proopiomelanocortin product secreted as a homodimer. *Mol. Endocrinol.* **1988**, *2* (11), 1108–14.
(45) Fenger, M.; Johnsen, A. H. Alpha-amidated peptides derived from pro-opiomelanocortin in human pituitary tumours. *J. Endocrinol.* **1988**, *118* (2), 329–38.
(46) Missler, M.; Sudhof, T. C. Neurexophilins form a conserved family of neuropeptide-like glycoproteins. *J. Neurosci.* **1998**, *18* (10), 3630–8.
(47) Mentlein, R. Proline residues in the maturation and degradation of peptide hormones and neuropeptides. *FEBS Lett.* **1988**, *234* (2), 251–6.
(48) Yoshihara, M.; Montana, E. S. The synaptotagmins: calcium sensors for vesicular trafficking. *Neuroscientist* **2004**, *10* (6), 566–74.
(49) Crambert, G.; Geering, K. FXYD proteins: new tissue-specific regulators of the ubiquitous Na,K-ATPase. *Sci. STKE* **2003**, *2003* (166), RE1.
(50) Feschenko, M. S.; Donnet, C.; Wetzel, R. K.; Asinovski, N. K.; Jones, L. R.; Sweadner, K. J. Phospholemman, a single-span membrane protein, is an accessory protein of Na,K-ATPase in cerebellum and choroid plexus. *J. Neurosci.* **2003**, *23* (6), 2161–9.
(51) Bhandari, V.; Palfree, R. G.; Bateman, A. Isolation and sequence of the granulin precursor cDNA from human bone marrow reveals tandem cysteine-rich granulin domains. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89* (5), 1715–9.
(52) Duguay, S. J.; Jin, Y.; Stein, J.; Duguay, A. N.; Gardner, P.; Steiner, D. F. Post-translational processing of the insulin-like growth factor-2 precursor. Analysis of O-glycosylation and endoproteolysis. *J. Biol. Chem.* **1998**, *273* (29), 18443–51.
(53) Rinderknecht, E.; Humbel, R. E. Polypeptides with nonsuppressible insulin-like and cell-growth promoting activities in human serum: isolation, chemical characterization, and some biological properties of forms I and II. *Proc. Natl. Acad. Sci. U.S.A.* **1976**, *73* (7), 2365–9.
(54) Tian, X. C.; Chen, M. J.; Pantschenko, A. G.; Yang, T. J.; Chen, T. T. Recombinant E-peptides of pro-IGF-I have mitogenic activity. *Endocrinology* **1999**, *140* (7), 3387–90.
(55) Higashiyama, S.; Lau, K.; Besner, G. E.; Abraham, J. A.; Klagsbrun, M. Structure of heparin-binding EGF-like growth factor. Multiple forms, primary structure, and glycosylation of the mature protein. *J. Biol. Chem.* **1992**, *267* (9), 6205–12.
(56) Miyamoto, S.; Yagi, H.; Yotsumoto, F.; Kawarabayashi, T.; Mekada, E. Heparin-binding epidermal growth factor-like growth factor as a novel targeting molecule for cancer therapy. *Cancer Sci.* **2006**, *97* (5), 341–7.
(57) Sakurai, T.; Amemiya, A.; Ishii, M.; Matsuzaki, I.; Chemelli, R. M.; Tanaka, H.; Williams, S. C.; Richardson, J. A.; Kozlowski, G. P.; Wilson, S.; Arch, J. R.; Buckingham, R. E.; Haynes, A. C.; Carr, S. A.; Annan, R. S.; McNulty, D. E.; Liu, W. S.; Terrett, J. A.; Elshourbagy, N. A.; Bergsma, D. J.; Yanagisawa, M. Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **1998**, *92* (4), 573–85.
(58) de Lecea, L.; Kilduff, T. S.; Peyron, C.; Gao, X.; Foye, P. E.; Danielson, P. E.; Fukuhara, C.; Battenberg, E. L.; Gautvik, V. T.; Bartlett, F. S.; Frankel, W. N.; van den Pol, A. N.; Bloom, F. E.; Gautvik, K. M.; Sutcliffe, J. G. The hypocretins: hypothalamus-specific peptides with neuroexcitatory activity. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95* (1), 322–7.
(59) Hungs, M.; Mignot, E. Hypocretin/orexin, sleep and narcolepsy. *Bioessays* **2001**, *23* (5), 397–408.
(60) Adachi, J.; Kumar, C.; Zhang, Y.; Olsen, J. V.; Mann, M. The human urinary proteome contains more than 1500 proteins including a large proportion of membranes proteins. *Genome Biol.* **2006**, *7* (9), R80.
(61) Pilch, B.; Mann, M., Large-scale and high-confidence proteomic analysis of human seminal plasma. *Genome Biol.* **2006**, *7* (5), R40.
(62) de Souza, G. A.; Godoy, L. M.; Mann, M. Identification of 491 proteins in the tear fluid proteome reveals a large number of proteases and protease inhibitors. *Genome Biol.* **2006**, *7* (8), R72.
(63) Olsen, J. V.; Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (37), 13417–22.
(64) Zeeberg, B. R.; Feng, W.; Wang, G.; Wang, M. D.; Fojo, A. T.; Sunshine, M.; Narasimhan, S.; Kane, D. W.; Reinhold, W. C.; Lababidi, S.; Bussey, K. J.; Riss, J.; Barrett, J. C.; Weinstein, J. N. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **2003**, *4* (4). R28.
(65) Harris, M. A.; Clark, J.; Ireland, A.; Lomax, J.; Ashburner, M.; Foulger, R.; Eilbeck, K.; Lewis, S.; Marshall, B.; Mungall, C.; Richter, J.; Rubin, G. M.; Blake, J. A.; Bult, C.; Dolan, M.; Drabkin, H.; Eppig, J. T.; Hill, D. P.; Ni, L.; Ringwald, M.; Balakrishnan, R.; Cherry, J. M.; Christie, K. R.; Costanzo, M. C.; Dwight, S. S.; Engel, S.; Fisk, D. G.; Hirschman, J. E.; Hong, E. L.; Nash, R. S.; Sethuraman, A.; Theesfeld, C. L.; Botstein, D.; Dolinski, K.; Feierbach, B.; Berardini, T.; Mundodi, S.; Rhee, S. Y.; Apweiler, R.; Barrell, D.; Camon, E.; Dimmer, E.; Lee, V.; Chisholm, R.; Gaudet, P.; Kibbe, W.; Kishore, R.; Schwarz, E. M.; Sternberg, P.; Gwinn, M.; Hannick, L.; Wortman, J.; Berriman, M.; Wood, V.; de la Cruz, N.; Tonellato, P.; Jaiswal, P.; Seigfried, T.; White, R. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **2004**, *32* (Database issue), D258–61.
(66) Pan, S.; Zhu, D.; Quinn, J. F.; Peskind, E. R.; Montine, T. J.; Lin, B.; Goodlett, D. R.; Taylor, G.; Eng, J.; Zhang, J. A combined dataset of human cerebrospinal fluid proteins identified by multi-

dimensional chromatography and tandem mass spectrometry. *Proteomics* **2007**, *7* (3), 469–73.
(67) Huhmer, A. F.; Biringer, R. G.; Amato, H.; Fonteh, A. N.; Harrington, M. G. Protein analysis in human cerebrospinal fluid: Physiological aspects, current progress and future challenges. *Dis. Markers* **2006**, *22* (1–2), 3–26.
(68) States, D. J.; Omenn, G. S.; Blackwell, T. W.; Fermin, D.; Eng, J.; Speicher, D. W.; Hanash, S. M. Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat. Biotechnol.* **2006**, *24* (3), 333–8.
(69) Blobel, C. P. Remarkable roles of proteolysis on and beyond the cell surface. *Curr. Opin. Cell Biol.* **2000**, *12* (5), 606–12.
(70) Pisitkun, T.; Shen, R. F.; Knepper, M. A. Identification and proteomic profiling of exosomes in human urine. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (36), 13368–73.
(71) Selkoe, D. J. Alzheimer's disease: genes, proteins, and therapy. *Physiol. Rev.* **2001**, *81* (2), 741–66.
(72) Haass, C.; De Strooper, B. The presenilins in Alzheimer's disease--proteolysis holds the key. *Science* **1999**, *286* (5441), 916–9.
(73) Neumann, S.; Schobel, S.; Jager, S.; Trautwein, A.; Haass, C.; Pietrzik, C. U.; Lichtenthaler, S. F. Amyloid precursor-like protein 1 influences endocytosis and proteolytic processing of the amyloid precursor protein. *J. Biol. Chem.* **2006**, *281* (11), 7583–94.
(74) Weber, B.; Schaper, C.; Scholz, J.; Bein, B.; Rodde, C.; Tonner, P. Interaction of the amyloid precursor like protein 1 with the alpha(2A)-adrenergic receptor increases agonist-mediated inhibition of adenylate cyclase. *Cell Signal* **2006**, *18* (10), 1748–1757.
(75) Roden, L. Structure and metabolism of connective tissue proteoglycans. In *The Biochemistry of Glycoproteins and Proteoglycans*; Lennarz, W., Ed.; Plenum Press: NY, 1980; pp 269–314.
(76) Kähnert, H.; Paddenberg, R.; Kleesiek, K. Simultaneous secretion of xylosyltransferase and chondroitin sulphate proteoglycane in chondrocyte cultures. *Eur. J. Clin. Chem. Clin. Biochem.* **1991**, *29*, 624–625.
(77) Kleesiek, K.; Reinards, R.; Okusi, J.; Wolf, B.; Greiling, H. UDP-D-xylose: proteoglycan core protein beta-D-xylosyltransferase: a new marker of cartilage destruction in chronic joint diseases. *J. Clin. Chem. Clin. Biochem.* **1987**, *25* (8), 473–81.
(78) Gotting, C.; Sollberg, S.; Kuhn, J.; Weilke, C.; Huerkamp, C.; Brinkmann, T.; Krieg, T.; Kleesiek, K. Serum xylosyltransferase: a new biochemical marker of the sclerotic process in systemic sclerosis. *J. Invest. Dermatol.* **1999**, *112* (6), 919–24.
(79) Ghiso, J.; Rostagno, A.; Tomidokoro, Y.; Lashley, T.; Bojsen-Moller, M.; Braendgaard, H.; Plant, G.; Holton, J.; Lal, R.; Revesz, T.; Frangione, B. Genetic alterations of the BRI2 gene: familial British and Danish dementias. *Brain Pathol.* **2006**, *16* (1), 71–9.
(80) Vidal, R.; Frangione, B.; Rostagno, A.; Mead, S.; Revesz, T.; Plant, G.; Ghiso, J. A stop-codon mutation in the BRI gene associated with familial British dementia. *Nature* **1999**, *399* (6738), 776–81.
(81) Vidal, R.; Revesz, T.; Rostagno, A.; Kim, E.; Holton, J. L.; Bek, T.; Bojsen-Moller, M.; Braendgaard, H.; Plant, G.; Ghiso, J.; Frangione, B. A decamer duplication in the 3′ region of the BRI gene originates an amyloid peptide that is associated with dementia in a Danish kindred. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97* (9), 4920–5.
(82) Fotinopoulou, A.; Tsachaki, M.; Vlavaki, M.; Poulopoulos, A.; Rostagno, A.; Frangione, B.; Ghiso, J.; Efthimiopoulos, S. BRI2 interacts with amyloid precursor protein (APP) and regulates amyloid beta (Abeta) production. *J. Biol. Chem.* **2005**, *280* (35), 30768–72.
(83) Paul, S.; Lombroso, P. J. Receptor and nonreceptor protein tyrosine phosphatases in the nervous system. *Cell. Mol. Life Sci.* **2003**, *60* (11), 2465–82.
(84) Harroch, S.; Furtado, G. C.; Brueck, W.; Rosenbluth, J.; Lafaille, J.; Chao, M.; Buxbaum, J. D.; Schlessinger, J. A critical role for the protein tyrosine phosphatase receptor type Z in functional recovery from demyelinating lesions. *Nat. Genet.* **2002**, *32* (3), 411–4.
(85) Muller, S.; Kunkel, P.; Lamszus, K.; Ulbricht, U.; Lorente, G. A.; Nelson, A. M.; von Schack, D.; Chin, D. J.; Lohr, S. C.; Westphal, M.; Melcher, T. A role for receptor tyrosine phosphatase zeta in glioma cell migration. *Oncogene* **2003**, *22* (43), 6661–8.
(86) Solimena, M.; Dirkx, R., Jr.; Hermel, J. M.; Pleasic-Williams, S.; Shapiro, J. A.; Caron, L.; Rabin, D. U. ICA 512, an autoantigen of type I diabetes, is an intrinsic membrane protein of neurosecretory granules. *EMBO J.* **1996**, *15* (9), 2102–14.
(87) Sun, Q. L.; Wang, J.; Bookman, R. J.; Bixby, J. L. Growth cone steering by receptor tyrosine phosphatase delta defines a distinct class of guidance cue. *Mol. Cell Neurosci.* **2000**, *16* (5), 686–95.
(88) Zhang, Y.; Adachi, J.; Olsen, J. V.; Shi, R.; de Souza, G.; Pasini, E.; Foster, L. J.; Macek, B.; Zougman, A.; Kumar, C.; Wisniewski, J. R.; Jun, W.; Mann, M., MAPU: Max-Planck Unified database of organellar, cellular, tissue and body fluid proteomes. *Nucleic Acids Res.* **2006**.

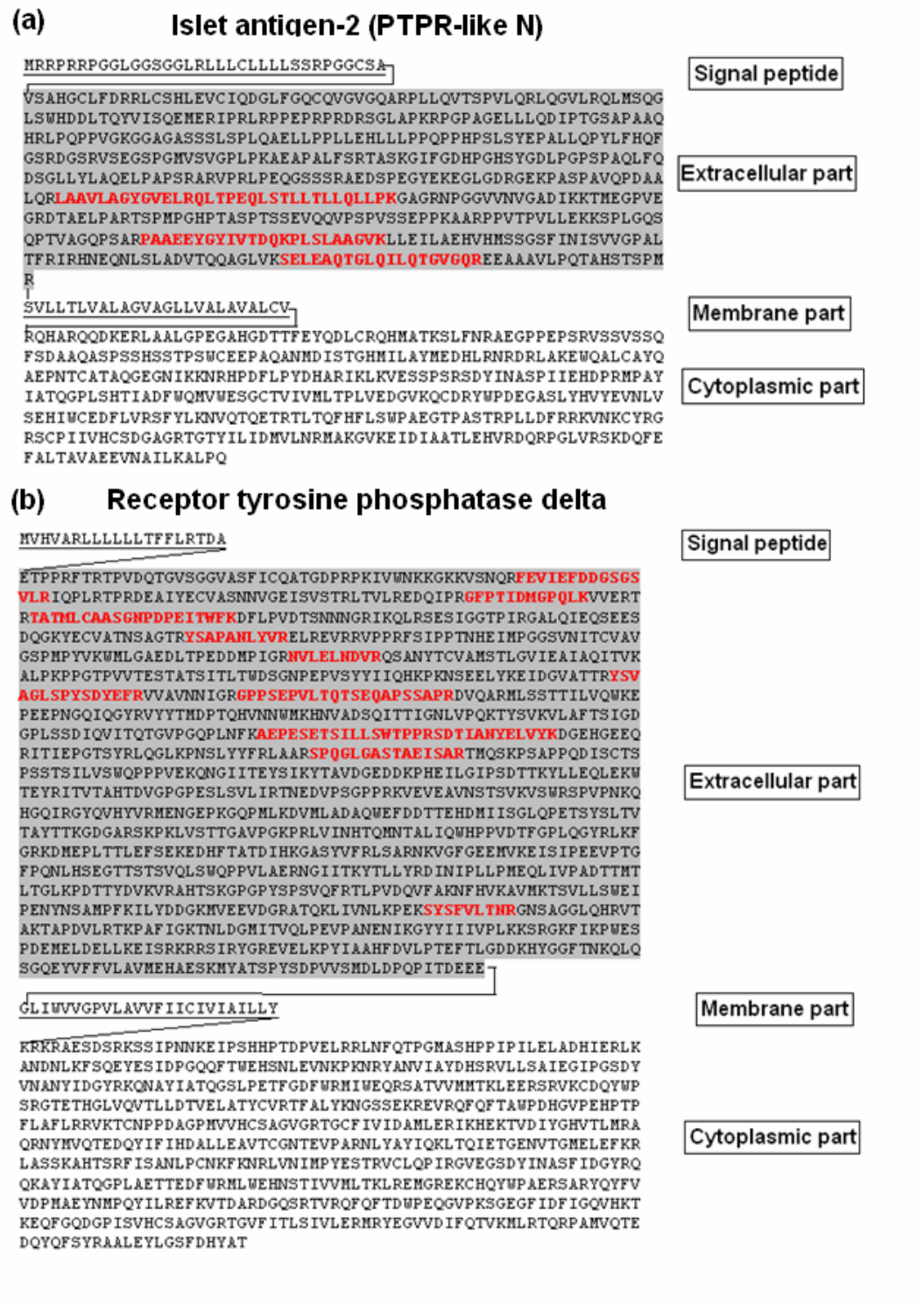
PR070501K

Supplementary Figure 1



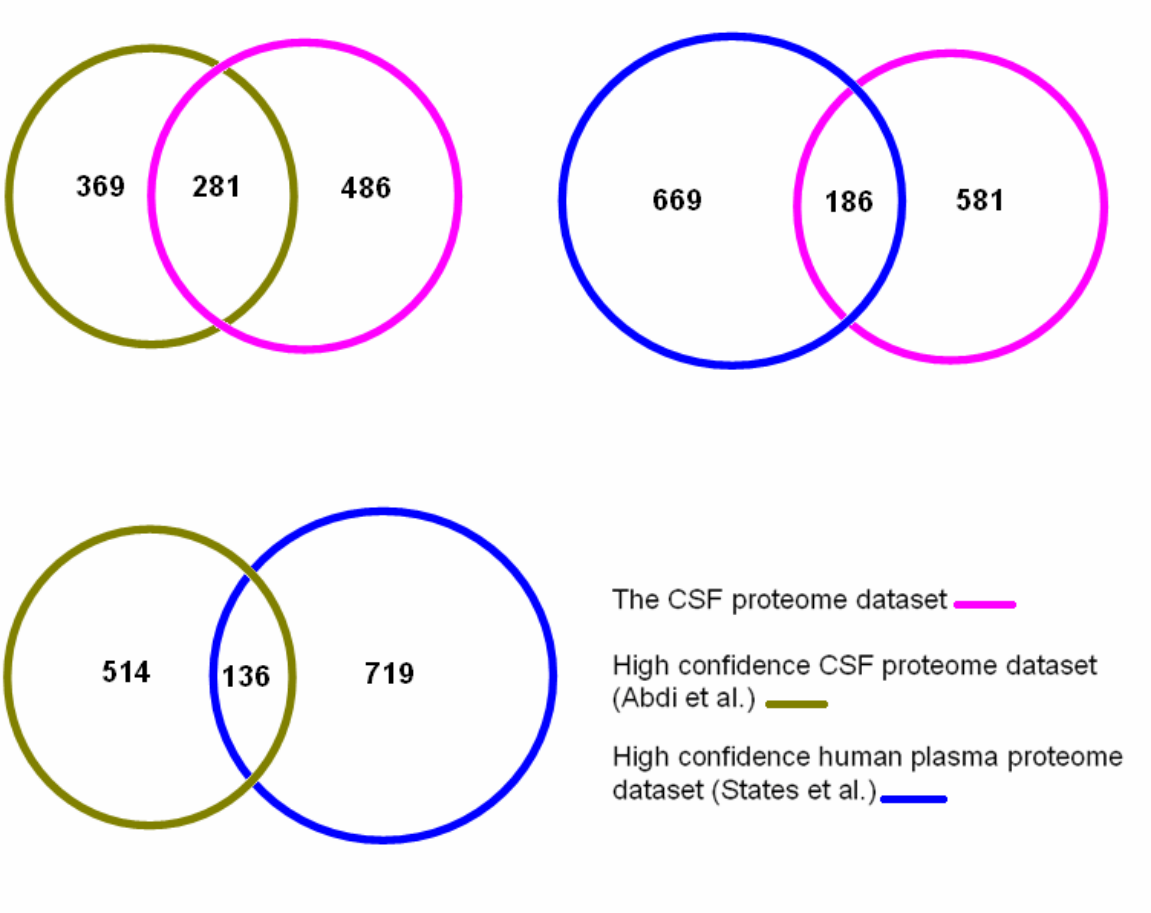
Supplementary Figure 2

Sequence coverage of *PTPR-like N* and *PTPR–delta* at the proteome level. Sequences identified during the CSF proteome mapping are shown in red.



Supplementary Figure 3

Comparison of CSF proteome with Human Plasma Proteome



Supplementary Figure 4

Sequence coverage of receptor tyrosine phosphatase zeta at the proteome and peptidome level. Sequences identified during the CSF proteome mapping are shown in red. The sequence identified during the CSF peptidome profiling is shown in blue.

Future directions

Neuropeptides are key signaling molecules in the synapses and represent a major part of the ligands of G protein-coupled receptors (GPCRs), the main targets of contemporary pharmaceuticals (*83*). About one hundred and twenty of non-sensory receptors identified as GPCRs from the Human Genome Project have no known endogenous interactors. In addition to that, little is known about the other, non-GPCR-type novel transmembrane receptors with names such as "KIAA…", "FAM…" or "Predicted…" Clearly identifying a cell surface target of the known endogenous ligand is easier than identifying an endogenous ligand for the transmembrane receptor. The latest developments in capillary separation science and improvements in sensitivity and accuracy of mass spectrometers have provided us with unique opportunities to probe the neuropeptidome to great depths and to identify novel low abundant neuropeptides (*54, 69*). Neuropeptides are harbored in synaptosomal vesicles. Synaptosomes are artificial membranous structures containing the components of nerve synapses created during the process of brain tissue homogenization (*84*). I believe that a simple enrichment of synaptosomal fraction with the following lysis of the synaptosomes to release the neuropeptides could significantly improve the sensitivity of the neuropeptidome profiling compared to the profiling of whole brain tissue homogenate. As a continuation of the neuropeptidome investigation, I already performed analysis of the rat brain synaptosomal neuropeptides by nanoLC-MS and discovered a number of novel neuropeptides. One of my goals is to characterize interactors/targets of some of the identified novel neuropeptides - I think that the outcome could be beneficial for our understanding of basic physiological and pathophysiological mechanisms and, potentially, lead to novel therapies.

Outlook

There is no doubt that further developments in high accuracy and high resolution mass spectrometry instrumentation will provide scientists with even faster and accurate discovery tools. At this moment the LTQ-Orbitrap justifiably dominates the proteomics market and it seems there is no viable competition for this unique instrument. Nevertheless, one cannot exclude, for example, that the ongoing evolution of the TOF instrumentation possibly will lead to creation of a new highly accurate mass spectrometer, e.g. an LTQ-TOF, which due to its nature should operate faster than the LTQ-Orbitrap and could provide similar mass accuracy and resolution. Then, perhaps, the time will come when proteomes are confidently characterized by utilizing highly accurate data at both MS and MS/MS levels. Much needed developments in the area of protein identification and characterization software will undoubtedly facilitate data analysis and, potentially, completely eliminate manual data analysis. Advances in sample preparation and capillary separation technologies will definitely increase the overall sensitivity of the LC-MS instrumentation. Hopefully, biologists will start to routinely use this powerful proteomics tool-box, which certainly will result in many exciting discoveries.

References

1. J. J. Thomson, *Philosophical Magazine* **44**, 293 (1897).
2. W. Stephens, *Physical Review* **69**, 691 (1946).
3. W. C. Wiley, I. H. McLaren, *Review of Scientific Instruments* **26**, 1150 (1955).
4. B. A. Mamyrin, V. I. Karataev, *Soviet Physiks JETP*, 374 (1973).
5. W. Paul, H. Steinwedel. Ger.Pat. 944 900 (1956); US Pat. 2,939,952 (1960).
6. G. Stafford, Jr., *J Am Soc Mass Spectrom* **13**, 589 (Jun, 2002).
7. E. O. Lawrence, M. S. Livingston, *Physical Review* **40**, 19 (1932).
8. M. B. Comisarow, A. G. Marshall, *J Mass Spectrom* **31**, 581 (Jun, 1996).
9. I. V. Chernushevich, A. V. Loboda, B. A. Thomson, *J Mass Spectrom* **36**, 849 (Aug, 2001).
10. K. H. Kingdon, *Physical Review* **21**, 408 (1923).
11. A. Makarov, *Anal Chem* **72**, 1156 (Mar 15, 2000).
12. M. S. B. Munson, F. H. Field, *J Am Chem Soc* **88**, 2621 (1966).
13. H. M. Fales *et al.*, *Recent Prog Horm Res* **28**, 591 (1972).
14. M. Yamashita, J. Fenn, *J Phys Chem* **88**, 4451 (1984).
15. J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, C. M. Whitehouse, *Science* **246**, 64 (Oct 6, 1989).
16. M. Karas, F. Hillenkamp, *Anal Chem* **60**, 2299 (Oct 15, 1988).
17. K. Tanaka, H. Waki, Y. Ido, *Rapid Commun. Mass Spectrom* **2**, 151 (1988).
18. R. C. Beavis, B. T. Chait, *Rapid Commun Mass Spectrom* **3**, 436 (Dec, 1989).
19. M. R. Emmett, R. M. Caprioli, *J Am Soc Mass Spectrom* **5**, 605 (1994).
20. O. Vorm, P. Roepstorff, M. Mann, *Anal Chem* **66**, 3281 (1994).
21. M. Wilm, M. Mann, *Anal Chem* **68**, 1 (Jan 1, 1996).
22. A. Shevchenko, M. Wilm, O. Vorm, M. Mann, *Anal Chem* **68**, 850 (Mar 1, 1996).
23. H. Steen, M. Mann, *Nat Rev Mol Cell Biol* **5**, 699 (Sep, 2004).
24. M. Golay, paper presented at the Gas Chromatography 1958 (Amsterdam Symposium), Amsterdam, 1958.
25. C. G. Horvath, B. A. Preiss, S. R. Lipsky, *Anal Chem* **39**, 1422 (Oct, 1967).
26. D. Ishii, *JASCO Report* **11**, 1 (1974).
27. J. Kirkland, J. De Stefano, *J. Chromatogr. Sci.* **8**, 309 (1970).
28. R. E. Majors, *Anal Chem* **44**, 1722 (1971).
29. R. D. Dandeneau, E. H. Zerenner, *J. High Resolut. Chromatogr. Chromatogr. Commun.* **2**, 351 (1979).
30. D. F. Hunt *et al.*, *Science* **255**, 1261 (Mar 6, 1992).
31. C. L. Gatlin, G. R. Kleemann, L. G. Hays, A. J. Link, J. R. Yates, 3rd, *Anal Biochem* **263**, 93 (Oct 1, 1998).
32. Y. Ishihama, J. Rappsilber, J. S. Andersen, M. Mann, *J Chromatogr A* **979**, 233 (Dec 6, 2002).
33. W. Haas *et al.*, *Mol Cell Proteomics* **5**, 1326 (Jul, 2006).
34. D. H. Lundgren, D. K. Han, J. K. Eng, *Curr Protoc Bioinformatics* **Chapter 13**, Unit 13 3 (Jul, 2005).
35. D. N. Perkins, D. J. Pappin, D. M. Creasy, J. S. Cottrell, *Electrophoresis* **20**, 3551 (Dec, 1999).
36. R. Craig, R. C. Beavis, *Bioinformatics* **20**, 1466 (Jun 12, 2004).
37. S. E. Ong, G. Mittler, M. Mann, *Nat Methods* **1**, 119 (Nov, 2004).
38. M. M. Savitski, M. L. Nielsen, R. A. Zubarev, *Mol Cell Proteomics* **5**, 935 (May, 2006).
39. B. Ma *et al.*, *Rapid Commun Mass Spectrom* **17**, 2337 (2003).
40. E. Pitzer, A. Masselot, J. Colinge, *Proteomics* **7**, 3051 (Sep, 2007).
41. J. A. Taylor, R. S. Johnson, *Anal Chem* **73**, 2594 (Jun 1, 2001).
42. M. I. Korsunskii, V. A. Basakutsa, *Sov. Physics-Tech. Phys.* **3**, 1396 (1958).

43. R. D. Knight, *Appl. Phys. Lett* **38**, 221 (1981).
44. L. N. Gall, Y. K. Golikov, M. L. Aleksandrov, Y. E. Pechalina, H. N.A. (1986).
45. M. Scigelova, A. Makarov, *Proteomics* **6 Suppl 2**, 16 (Sep, 2006).
46. A. A. Makarov. US Pat. 5,886,346 (1999).
47. P. Oksman, *Int. J. Mass Spectrom. Ion Processes* **141**, 67 (1995).
48. J. V. Olsen *et al.*, *Mol Cell Proteomics* **4**, 2010 (Dec, 2005).
49. D. J. Douglas, A. J. Frank, D. Mao, *Mass Spectrom Rev* **24**, 1 (Jan-Feb, 2005).
50. R. A. Zubarev, Hakansson, P., Sundquist, B.,, *Anal. Chem.* **68**, 4060 (1996).
51. J. R. Yates, D. Cociorva, L. Liao, V. Zabrouskov, *Anal Chem* **78**, 493 (Jan 15, 2006).
52. A. Makarov *et al.*, *Anal Chem* **78**, 2113 (Apr 1, 2006).
53. J. V. Olsen *et al.*, *Nat Methods* **4**, 709 (Sep, 2007).
54. A. Zougman *et al.*, *J Proteome Res* **7**, 386 (Jan, 2008).
55. A. Zougman, P. Ziolkowski, M. Mann, J. R. Wisniewski, *Curr Biol* **18**, 1760 (Nov 25, 2008).
56. E. Birney *et al.*, *Nucleic Acids Res* **34**, D556 (Jan 1, 2006).
57. D. Maglott, J. Ostell, K. D. Pruitt, T. Tatusova, *Nucleic Acids Res* **35**, D26 (Jan, 2007).
58. P. Bork, *Genome Res* **10**, 398 (Apr, 2000).
59. R. Guigo *et al.*, *Genome Biol* **7 Suppl 1**, S2 1 (2006).
60. S. Tanner *et al.*, *Genome Res* **17**, 231 (Feb, 2007).
61. D. Fermin *et al.*, *Genome Biol* **7**, R35 (2006).
62. F. Desiere *et al.*, *Genome Biol* **6**, R9 (2005).
63. V. E. US, J. H. Gaddum, *J Physiol* **72**, 74 (Jun 6, 1931).
64. M. M. Chang, S. E. Leeman, *J Biol Chem* **245**, 4784 (Sep 25, 1970).
65. T. Hokfelt, T. Bartfai, F. Bloom, *Lancet Neurol* **2**, 463 (Aug, 2003).
66. B. R. Southey, A. Amare, T. A. Zimmerman, S. L. Rodriguez-Zas, J. V. Sweedler, *Nucleic Acids Res* **34**, W267 (Jul 1, 2006).
67. F. Bergeron, R. Leduc, R. Day, *J Mol Endocrinol* **24**, 1 (Feb, 2000).
68. B. A. Eipper, S. L. Milgram, E. J. Husten, H. Y. Yun, R. E. Mains, *Protein Sci* **2**, 489 (Apr, 1993).
69. M. Svensson *et al.*, *Anal Chem* **79**, 15 (Jan 1, 2007).
70. B. Kuster, M. Schirle, P. Mallick, R. Aebersold, *Nat Rev Mol Cell Biol* **6**, 577 (Jul, 2005).
71. A. Zougman, Ziółkowski, P., Mann, M. and Wiśniewski J.R. , *Current Biology* (2008).
72. A. Zougman, J. R. Wisniewski, *J Proteome Res* **5**, 925 (Apr, 2006).
73. J. R. Wisniewski, A. Zougman, S. Kruger, M. Mann, *Mol Cell Proteomics* **6**, 72 (Jan, 2007).
74. J. M. Gott, R. B. Emeson, *Annu Rev Genet* **34**, 499 (2000).
75. E. J. Thompson, G. Keir, *Ann Clin Biochem* **27 (Pt 5)**, 425 (Sep, 1990).
76. G. J. Siegel, B. W. Agranoff, R. W. Albers, S. K. Fisher , M. D. Uhler, Eds., *Basic Neurochemistry - Molecular, Cellular, and Medical Aspects* (1999).
77. D. A. Seehusen, M. M. Reeves, D. A. Fomin, *Am Fam Physician* **68**, 1103 (Sep 15, 2003).
78. B. N. Hammack *et al.*, *Mult Scler* **10**, 245 (Jun, 2004).
79. F. S. Berven, K. Flikka, M. Berle, C. Vedeler, R. J. Ulvik, *Curr Pharm Biotechnol* **7**, 147 (Jun, 2006).
80. M. Puchades *et al.*, *Brain Res Mol Brain Res* **118**, 140 (Oct 21, 2003).
81. F. Abdi *et al.*, *J Alzheimers Dis* **9**, 293 (Aug, 2006).
82. J. Zhang *et al.*, *Neurobiol Aging* **26**, 207 (Feb, 2005).
83. S. Chung, O. Civelli, *Neuropeptides* **40**, 233 (Aug, 2006).
84. V. P. Whittaker, *J Neurocytol* **22**, 735 (Sep, 1993).

Acknowledgements

I believe that during the past years I have been very lucky to have met extraordinary individuals, physicians and scientists whose ideas and attitudes greatly influenced my personal development.

I am thankful to late Prof. Stanislav Shipov, M.D. of the 1st Moscow Municipal Hospital, late Prof. Alexander Grinberg, M.D., Ph.D., D.Sc. of the 15th Moscow Municipal Hospital, Prof. Olga Shevchenko, M.D., Ph.D., D.Sc. of the Moscow Institute of Transplantology and Artificial Organs, Anna Chernova, M.D., Ph.D. of the Moscow Institute of Transplantology and Artificial Organs, Prof. Yoav Dickstein, Ph.D. of the Leumit Central Lab, late Cemal Kuyas, Ph.D., "The Master and Commander" of MDS Proteomics Inc.

I am grateful to Valentina Bikeeva, M.D. of the 15th Moscow Municipal Hospital who taught me the real value of life.

I am thankful to Paul Taylor, M.D. of the University of Toronto who introduced me to the exciting world of biological mass spectrometry.

I am indebted to Prof. Matthias Mann, Ph.D. and Prof. Jacek Wiśniewski, Ph.D. of the Max Planck Institute for Biochemistry for generous and caring advice, support and guidance.

I also would like to thank the fellow group members at the Department of the Proteomics and Signal Transduction of the Max Planck Institute for Biochemistry for their kind help.

As well I want to express gratitude to Vadim Gorodetskiy, M.D., Ph.D., Pavel Metalnikov, M.Sc., Adrian Pasculescu, Ph.D. for their support and friendship.

And, of course, this work could not be made possible without the endless support of my family, my mother Elena and my father Emanuil, my wife Inga and daughter Uma.

Thank you.

Curriculum Vitae

Name: Alexandre Zougman

Gender: male

Date of birth: July 14, 1971

Nationality: Canada, Israel, Russia

Family status: married

E-mail: zougman@gmail.com

Professional Profile

- ✓ *In-depth* knowledge of biological mass spectrometry and hyphenated techniques
- ✓ *Thorough* knowledge of protein biochemistry, molecular biology, cell imaging and bioinformatics
- ✓ 3 years as a scientist in one of the world's leading proteomics labs; 5+ years of industrial hands-on experience in mass spectrometry-based proteomics
- ✓ Excellent communication and interpersonal skills
- ✓ Started and led successful discovery program efforts
- ✓ Committed to highest quality creative work

Education

M.D., M.Sc. Russian State Medical University, Moscow, (1994)

Professional Experience

2006 - present Scientist, Prof. Matthias Mann's Lab, Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Martinsried, Germany

- Biological mass spectrometry, protein biochemistry, molecular biology, cell imaging, bioinformatics. Discovery and characterization of extended nuclear protein forms resulting from a novel mRNA processing mechanism. Application of advanced proteomics techniques towards identifying and characterizing novel neuropeptides and their targets. Characterization of putative post-translational modifications of nuclear proteins. Quantitative proteomics. Development and optimization of sample preparation methods for biomarker discovery.

2000 - 2005 Senior Scientist, Proteomics Group, MDS Proteomics Inc., Toronto, Canada

- Development and implementation of methodologies in support of the proteomics projects for

external clients. Supervision and training of junior scientists.

1999 - 2000 Mass Spectrometrist, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada

- Mastering the basics of biological mass spectrometry. Analytical techniques in proteomics research.

1996 - 1999 Clinical Chemist, Senior Staff Member, Central Diagnostic Lab, Leumit Health Fund, Israel

- Optimization and implementation of immunochemical tools in diagnostics of hormonal and malignant disorders.

1993 - 1995 Research Associate, Division of Immunochemistry and Biomaterials, Institute of Transplantology and Artificial Organs, Russian Academy of Medical Sciences, Russia

- Mastering protein biochemistry and immunochemistry. Biochemical and immunochemical characterization of leukocytic antigens.

Publications

1. **Zougman A**, Ziółkowski P, Mann M & Wiśniewski JR**.**
Evidence for Insertional RNA Editing in Humans. *Curr Biol*. 2008 Nov25;18(22):1760-5.

2. Wiśniewski JR, **Zougman A**, Nagaraj N, and Mann M**.**
Universal sample preparation method for proteome analysis combines the advantages of in-solution and in-gel digestion *(accepted for publication in Nature Methods).*

3. Lu A**, Zougman A**, Pudełko M, Bębenek M, Ziółkowski P, Mann M & Wiśniewski JR**.**
Mapping of Lysine Monomethylation of Linker Histones in Human Breast Cancer *(submitted to Journal of Proteome Research).*

4. Wiśniewski JR, **Zougman A**, Krüger S, Ziółkowski P, Pudełko M, Bębenek M, Mann M.
Constitutive and dynamic phosphorylation and acetylation sites on NUCKS, a hypermodified nuclear protein, studied by quantitative proteomics. *Proteins*. 2008 May 19. (Epub ahead of print).

5. Wiśniewski JR**, Zougman A**, Mann M.
N-epsilon-formylation of lysine is a widespread post-translational modification of nuclear proteins occurring at residues involved in regulation of chromatin function. *Nucleic Acids Res*. 2008 Feb;36(2):570-7.

6. **Zougman A**, Pilch B, Podtelejnikov A, Kiehntopf M, Schnabel C, Kumar C, Mann M.
Integrated analysis of the cerebrospinal fluid peptidome and proteome. *J Proteome Res.* 2008 Jan;7(1):386-99.

7. Shi R, Kumar C**, Zougman A**, Zhang Y, Podtelejnikov A, Cox J, Wiśniewski JR, Mann M.
Analysis of the mouse liver proteome using advanced mass spectrometry. *J Proteome Res.* 2007 Aug;6(8):2963-72.

8. Khanna R, **Zougman A**, Stanley EF.
A proteomic screen for presynaptic terminal N-type calcium channel (CaV2.2) binding partners. *J Biochem Mol Biol.* 2007 May 31;40(3):302-14.

9. Zhang Y, Zhang Y, Adachi J, Olsen JV, Shi R, de Souza G, Pasini E, Foster LJ, Macek B, **Zougman A**, Kumar C, Wiśniewski JR, Jun W, Mann M.
MAPU: Max-Planck Unified database of organellar, cellular, tissue and body fluid proteomes. *Nucleic Acids Res.* 2007 Jan;35 (Database issue):D771-9.

10. Wiśniewski JR, **Zougman A**, Krüger S, Mann M.
Mass spectrometric mapping of linker histone H1 variants reveals multiple acetylations, methylations, and phosphorylation as well as differences between cell culture and tissue. *Mol Cell Proteomics.* 2007 Jan;6(1):72-87.

11. **Zougman A**, Wiśniewski JR.
Beyond linker histones and high mobility group proteins: global profiling of perchloric acid soluble proteins. *J Proteome Res.* 2006 Apr;5(4):925-34.

12. Jin J, Smith FD, Stark C, Wells CD, Fawcett JP, Kulkarni S, Metalnikov P, O'Donnell P, Taylor P, Taylor L, **Zougman A**, Woodgett JR, Langeberg LK, Scott JD, Pawson T.
Proteomic, functional, and domain-based analysis of in vivo 14-3-3 binding proteins involved in cytoskeletal regulation and cellular organization. *Curr Biol.* 2004 Aug 24;14(16):1436-50.

13. Nelson B, Kurischko C, Horecka J, Mody M, Nair P, Pratt L, **Zougman A**, McBroom LD, Hughes TR, Boone C, Luca FC.
RAM: a conserved signaling network that regulates Ace2p transcriptional activity and polarized morphogenesis. *Mol Biol Cell.* 2003 Sep;14(9):3782-803.