Available online at www.sciencedirect.com
ScienceDirect
International Journal of Mass Spectrometry 259 (2007) 16–31
International Journal of Mass Spectrometry
www.elsevier.com/locate/ijms

Glycoproteomics: Past, present and future

Howard R. Morris [a,b,*], Sara Chalabi [a], Maria Panico [a], Mark Sutton-Smith [a], Gary F. Clark [c], David Goldberg [d], Anne Dell [a]

[a] *Division of Molecular Biosciences, Faculty of Natural Sciences, South Kensington Campus, Biochemistry Building, Imperial College, London SW7 2AZ, UK*
[b] *M-Scan Ltd., 3 Millars Business Centre, Fishponds Close, Wokingham RG41 2TZ, UK*
[c] *Department of Obstetrics and Gynecology, University of Wisconsin Medical School, Madison, WI 53792-6177, USA*
[d] *Palo Alto Research Center, Palo Alto, CA 94301, USA*

Received 5 July 2006; received in revised form 30 August 2006; accepted 1 September 2006
Available online 10 October 2006

Abstract

This invited paper charts the origins and progress of glycoproteomics mass spectrometry research at Imperial College, and in celebration of Donald Hunt's 65th birthday it puts into perspective some of the scientific influence which each group has had on the other over a period of some 35 years. We then describe and illustrate current nano-LC–ES–MS and MS/MS strategies for the structural assignment of N-linked glycosylation in proteins involved in sperm/egg fertilisation. Finally, we present recent progress in the automated interpretation of these glycopeptide data sets, which promises to supersede manual interpretation for many applications.

Keywords: Glycoproteomics; Mass spectrometry; Peptide sequencing; Carbohydrate sequencing; ZP3

1. Historical background to glycoproteomics at imperial

Donald Hunt's interaction with the Imperial College Biomolecular Mass Spectrometry Group began in the early 1970s when he attended a Cambridge University Chemical Laboratory seminar in the Williams group given by one of us (HRM) on the latest MS strategies developed for sequencing proteins of unknown structure following on from the original protein MS work on silk fibroin [1,2]. The seminar dealt with the idea and practical application of sequencing mixtures of enzymically derived peptides created by tryptic and/or other digestions of proteins, thus obviating the then necessary and rate-limiting step of purifying individual peptides for either classical Edman or MS sequencing. The seminar demonstrated the relative ease with which a dozen or so peptides of unknown structure in a mixture could be simultaneously sequenced from the map of molecular and fragment ions produced by fractional distillation from a sample probe using electron impact mass spectrometry [3–5]. This methodology required the initial derivatisation of the peptide mixtures to make them volatile and the breakthrough in applying the strategy was the discovery in 1972 that the permethylation chemistry used prior to that date had in fact been "blocking" the ability to observe many of the peptides in a mixture, particularly any containing Arg, His, Cys or Met, by creating involatile quaternary ammonium or sulphonium salts. The permethylation chemistry applied to peptides had in fact been borrowed from the carbohydrate field some years earlier, where reaction times of many hours or even days were recommended [6]. This had been reduced to 1 h for convenience in the peptide and sugar fields, and although salt formation had been suspected to be the reason why peptides containing the above amino acids had rarely been observed, the problem had defeated many groups' attempts to overcome it despite the use of several ingenious derivatisation strategies.

The solution, which Don appreciated as a chemist himself, came from an examination of the actual rate of the Hakomori permethylation reaction using dimsyl sodium anion and methyl iodide. In stable isotope dilution experiments we found surprisingly that the chemistry was essentially complete just 1 min after the addition of the alkylating agent, and furthermore that if the

* Corresponding author at: Division of Molecular Biosciences, Faculty of Natural Sciences, South Kensington Campus, Biochemistry Building, Imperial College, London SW7 2AZ, UK. Tel.: +44 2075945221; fax: +44 2072250458.
E-mail address: h.morris@imperial.ac.uk (H.R. Morris).

1387-3806/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijms.2006.09.002

reaction time was restricted to that order, then suddenly peptides containing the "impossible" amino acids were visible in the mass spectra created. It was concluded that salt formation was therefore a slower process, and that by incorporating a short methylation step, for the first time in 1972 [7,8] we had a global strategy for sequencing any protein-derived peptides by mass spectrometry.

Over the following decade, initially in the Cambridge Chemical and the MRC Molecular Biology laboratories, and after 1975 at Imperial College, London, our group applied the mixture mapping strategy to numerous biologically important peptide sequencing problems including the sequencing of complete proteins such as ribitol dehydrogenase [9], chloramphenicol transacetylase [10,11] and dihydrofolate reductase [12–14], and in difficult natural product analyses such as the structure elucidation of the first endogenous opiate "endorphin" which we christened enkephalin [15].

Foremost amongst the strengths of MS methodology at the time (and this is still true and relevant today) was the ability to detect unusual structural features or structural modifications, which differ in mass. In a particularly exciting collaboration with the Danish group under Staffan Magnusson a post-translational modification of the blood coagulation zymogen prothrombin was shown, by the MS strategies defined above, to involve the gamma-carboxylation of 10 specific glutamic acid residues in the N-terminal domain of the protein [16–18], of crucial significance to calcium and phospholipid binding at the site of a wound. We christened this new amino acid GLA. This study was extended to the discovery of other GLA-containing proteins including Factor X [19], shown to contain 12 specific carboxyl modifications, again in the N-terminal domain.

A biologically important and quite common post-translational modification of proteins is of course glycosylation, and our group started to study simple carbohydrates in the early 1970s in Cambridge, initially attempting to distinguish between isobaric sugars such as mannose and galactose in di- or trisaccharides by forming cyclic phenyl boronate derivatives prior to MS study. Another open seminar on this subject led to a very interesting collaboration with Robert Feeney of UC Davis in California and to our first introduction to the complex problems of glycoprotein characterisation. Feeney was visiting the Scott Polar Research Institute adjacent to the University Chemical Laboratory in Cambridge and was studying the structure of a fascinating group of Antarctic fish blood proteins which he had shown had potent antifreeze properties. He explained that the proteins were very difficult to sequence classically by Edman degradation, ironically because of the simplicity of their amino acid composition and the repetitivity of the sequence. Furthermore they were known to be glycosylated but the degree of modification and the carbohydrate structures were at that time also unknown. The carbohydrate was expected, however, to be bonded via an O-link to threonine because of the high proportion of THR in the amino acid analysis. Feeney had sufficient quantities of one of the proteins – a proline-containing column fraction coded antifreeze 8 or AF8 – for us to start to develop approaches to the structural characterisation. As we were to learn, problems of this type are far from trivial, even with today's techniques, not least because in addition to the problem of deriving amino acid sequence, there is the need to define carbohydrate composition (to distinguish the presence of isobaric sugars), the carbohydrate sequence, the peptide/carbohydrate linkage position(s), the carbohydrate/carbohydrate linkages and the configurations of linkages.

This first glycoprotein MS study proved to be seminal in demonstrating the power and flexibility of the MS approach, and its many advantages over classical methodology. In brief, from the same sample preparation using acetylation to block amino functions followed by the new short permethylation, using EI-MS it proved possible to determine the amino acid sequence of AF8 and demonstrate heterogeneity within it, to show unambiguously the presence of multiple positions of O-linked sugar attachment, and using CI-MS to show intense quasimolecular ions, to conclude that a single type of disaccharide, a hexosyl acetamidohexose, originally substituted each threonine in the structure prior to derivative formation (Fig. 1). Interestingly, the short Hakomori methylation developed for peptide analysis earlier, had here effectively labelled the carbohydrate substitution positions by beta elimination to leave dehydroalanine (a different mass to threonine) wherever carbohydrate had previously been linked. In a different experiment using NaOH elimination from the polypeptide backbone, followed by trimethylsilylation and MS analysis we were able to show by comparison with disaccharide standards of 1–4 linkage (*N*-acetyl lactosamine) and 1–3 linkage (lacto-*N*-biose I) that two key signals in the EI-MS spectra of AF8 (*m/z* 637 and 652) defined the linkage as 1–3 in this natural product (Fig. 2). Together with classical composition analysis the structure was thus defined as galactosyl 1–3 *N*-acetyl galactosamine [20]. This work was progressing at the same time as a novel high mass spectrometer using new high field magnet technology that we had commissioned for Imperial College was being built [21,22]. The typical mass range even for high resolution double focusing mass spectrometers at that time was less than 1000 making studies of larger molecules quite difficult. In the early 1970s we had done some basic calculations using the Atlas of Protein Structure and theoretically digesting with specific enzymes such as trypsin to determine an optimal mass range which would encompass most tryptic peptides at reasonable resolution and therefore mass accuracy. The answer was that we discovered that the majority of tryptic peptides fall below 3500 Da and in 1974 we managed to obtain grant funding to commission an instrument based on a high field magnet to deliver this specification at full accelerating voltage and thus full sensitivity. We then had the opportunity to study a number of projects which had previously been hampered by limited mass range. In particular, combining this equipment with a field desorption ion source which HRM had designed earlier in Cambridge, we were then able to define accurate nominal mass molecular ions on quite large molecules (>3000 amu) for the first time. Applying this new technology to the AF8 problem allowed good confirmation that the peptide length of the heterogeneous mixture observed was 14 residues, previously only inferred from a combination of N- and C-terminal sequence ions. Despite this success, our data remained unpublished except for a congress proceedings [23] because Prof. Feeney's group became aware

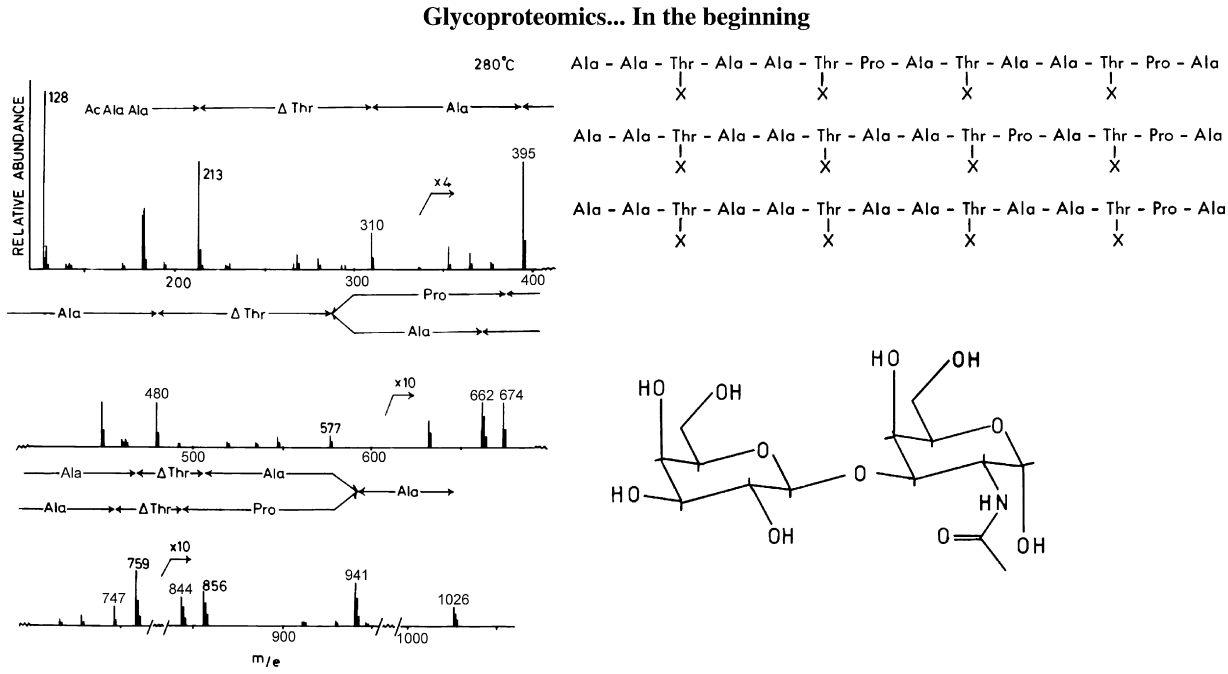

Fig. 1. EI mass spectrum and structural conclusions from the first glycopeptide MS study on antifreeze AF8 from the Antarctic fish *T. borchgevinki* [20,23].

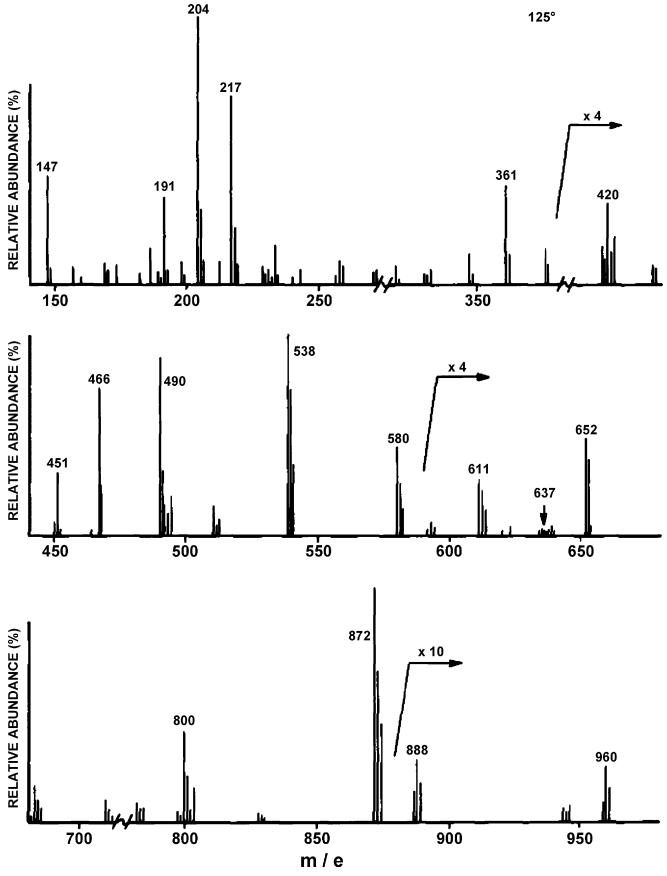

Fig. 2. EI mass spectrum of the trimethylsilyl derivative of the beta eliminated carbohydrate from AF8 showing the high 652:637 abundance ratio used to assign a 1–3 linkage in the hexosyl acetamidohexose structure determined by CI-MS [20,23].

of an NMR study of AF8 glycopeptides in which a galactosyl 1–4 *N*-acetyl galactosamine structure had been assigned to the carbohydrate [24]. Our main paper remained unsubmitted until we had the opportunity to have our proposed 1–3 structure synthesised and to prove that the mass spectrometric TMS data on the beta eliminated AF8 carbohydrate was identical to the 1–3 synthetic disaccharide. This must be one of the few cases where a carbohydrate linkage was misassigned by NMR and corrected by MS [20]!

Shortly after the AF8 paper was submitted for publication, HRM received an invitation from Donald Hunt to accept a Visiting Professorship at the University of Virginia for 6 months, to train some of his research students in biopolymer mass spectrometry. This was a wonderful trip in the spring and summer of 1978 to a great university and the interaction and scientific exchange with people in addition to Donald, such as Jeff Shabanowitz, Don's key man on the mass spectrometers, and Alex Buko and other bright students, was particularly rewarding.

On HRM's return to the UK, the AF8 success encouraged the group to attack a number of other difficult glycoprotein problems, including an early foray into the N-linked glycosylation analysis of prothrombin [25]. Shortly after this, however, the whole scenario of biopolymer analysis became much simpler with the invention of fast atom bombardment [26], the first soft ionisation method applicable to non-volatile materials which did not require the "black box" know-how of field desorption, or the use of potentially dangerous radioisotopes such as present in californium ionisation. The new FAB method allowed facile molecular weight determination and therefore the easier application of our mixture analysis strategy to the mapping of peptide mixtures [27], together with the provision of some sequence fragmentation to allow structure determination of unknowns [28]. With the high field magnet mass spectrometer now recognised as a

valuable instrument for gaining isotopically resolved and accurately calibrated unambiguous data well beyond the 3 kDa mass range, Don Hunt accepted the opportunity to spend some time in London when HRM invited him as Visiting Professor to Imperial in 1981–1982. With his background in quadrupoles and ours in magnetic sectors we had many interesting discusssions about the relative merits of the two very different types of instrumentation. Mass range and resolution were the crucial weaknesses of quadrupoles, and magnetic sectors were to dominate biopolymer analysis for almost another decade until the advent of the creation of multiply charged ion sources by electrospray by Yamashita and Fenn [29], and Alexandrov et al. [30], reducing the mass range requirement of the analyser as a function of charge state, and of course until the subsequent development of commercial instrumentation.

Utilising the new ease of FAB analysis which had the added advantage of increased sensitivity – into the low picomole range especially when incorporating the small M-Scan ion gun to give increased flux density of bombarding particles near the target [31] – numerous glycoprotein problems requiring a high degree of sensitivity could now be tackled more seriously. Glycoproteomic problems at the structural level divide themselves into O-linked glycosylation, normally at Ser or Thr in which the modification is usually of modest size (1–8 residues), as seen in the AF8 study, and N-linked glycosylation on asparagine residues in which larger modifications of so-called complex, high mannose or hybrid structures are attached [32]. Each type poses its own analytical problems and there is no single universal glycoproteomics mass spectrometric strategy for structural solutions. Having said that, the discovery and assignment of glycosylation is normally based on a mass shift observed in the quasimolecular ion from that expected for an unmodified peptide to some combination of the peptide mass with sugar increments. The discovery and interpretation of that mass shift can be and is made in many ways depending on the data set and experimental strategy. This can involve theoretical sugar mass additions to the peptide mass followed by searching of the data set for possible signals. This works well for O-linked discovery, and sometimes even for N-linked where a "trial structure" of a common bi-antennary addition, for example, will sometimes produce a hit, or at least take you into the region of a complex mass spectrum where signals of interest are located. Alternatively, searching for sugar mass differences or low mass ions caused by partial in-source fragmentation in MS spectra or induced by CAD in MS/MS spectra will allow you to lock-on to glycopeptide signals in a protein digest. Each type of study may present its own difficulties; there is no known consensus sequence at the peptide level for *O*-glycosylation, although it is recognised that we tend to observe attachment in areas of the sequence which are rich in serine and/or threonine residues together with proline. Location of the precise site(s) of *O*-glycosylation is therefore difficult where there may be multiple Ser/Thr candidates in the peptide sequence, and unless this problem can be solved by judicious choice of digest strategy, then the MS/MS fragmentation data will be crucial to a successful assignment. This in itself poses special problems because the internal energy required to break glycosidic bonds is less than that required to cleave peptide bonds, and careful tuning of collision energy and gas pressure is usually required to preserve sugar attachment data in the induced peptide fragmentation. In contrast to O-linked, there is a consensus sequence in the peptide structure for *N*-glycosylation (Asn-X-Ser/Thr where X is not Pro), making it easy to predict possible *N*-glycosylation sites, and the biosynthetic rules governing *N*-glycan synthesis can assist in reducing the thousands of possible structures which would derive from a consideration of mass alone. Nevertheless, the sheer size and often the heterogeneity of N-linked glycopeptide structures can make their assignment non-trivial, and later we describe current strategies and future research aimed at alleviating this problem. Collaborative successes using the fast atom bombardment/high field magnet combination in this period ranged from intact glycopeptide studies such as the discovery and characterisation of *O*-glycosylation on human Interleukin II [33] and recombinant erythropoietin [34], to released carbohydrate analyses such as the definition of N-linked structures on erythropoietin [35], Bowes melanoma tissue plasminogen activator [36] and proopiomelanocortin [37]. The advent of electrospray ionisation [29,30] allowed simplified on-line LCMS experiments in which heterogeneous populations of site-specific *N*-glycopeptides could be separated and identified more easily, eluting over particular regions of the chromatogram, as exemplified by research on the discovery and characterisation of the gender-specific glycosylation of glycodelin [38,39].

Most post-translational problems of the 1970s and 1980s, including the new strategies for S–S bridge analysis [40], relied on "mapping" the masses of relevant peptide/glycopeptide digests along the mass axis of the spectrum. This could involve the study of intact protein/glycoprotein digests or of discrete semi-purified LC "pools". It rarely involved the necessity for absolute purification, and the deliberate by-passing of this rate-limiting step by early mixture analysis strategies of the 1960s and 1970s was now extended using the new soft ionisation methods of FAB, electrospray and MALDI for both peptide/glycopeptide mixture mapping and released glycan mapping in the studies described above. There was, however, an important and increasing need, determined by biological availability, for higher sensitivity in these methodologies—to move the target sensitivity from the picomole into the femtomole range whilst preserving good mass accuracy. In order to achieve this, in 1987 the group proposed the building of a new "wide angle" focal plane MCP detector double focusing mass spectrometer to succeed the "low angle" (4%) experimental MS50 design [41] which required far too many step-jumps to cover a practicable mass range for real problems. The theoretical considerations and available MCP detector technology suggested that a much greater coverage could be achieved with a single "shot" using a focal plane geometry, making a true high sensitivity experiment feasible. The major manufacturers were however initially very reluctant to build a new true focal plane instrument, at what we felt was a reasonable cost, for a possibly very small market, and instead we had to compromise with a modification of the existing ZAB 2SE geometry to fulfill our specifications of (a) wide angle focal plane MCP detection at moderate (>1000 resolution in the survey mode and (b) concomitant selection of signals (low angle)

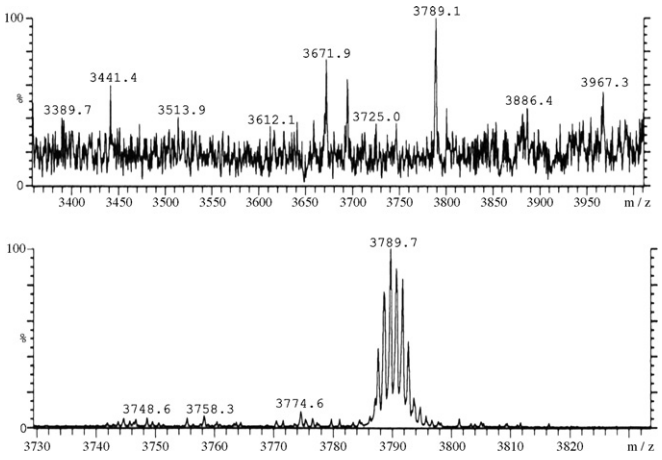

Fig. 3. A MALDI spectrum of calcitonin gene related peptide created on the dual focal plane detector ZAB2SE2FPD instrument showing the wide angle low resolution survey scan (top trace) from which signals of interest were switched instantaneously to the narrow angle high resolution array, allowing the study of biomolecules down to 10 fm with a mass accuracy of 0.1 Da. The data created from this instrument at the beginning of the 1990s showed us that the theoretical benefit of non-scanning MCP detection of up to 100-fold was realisable at the practical level [42].

for high resolution (10,000) MCP analysis to give good mass accuracy of 0.1 Da and isotopic resolution beyond 3000 amu. The ZAB2SE 2FPD was built in 1990 and immediately fulfilled its theoretical promise for non-scanning detection by allowing an increase in detection limits of important biomolecules of up to a factor of 100, down to the 10 fmol level, with an unsurpassed signal quality and isotopic resolution as demonstrated in work presented to the American Society of Mass Spectrometry in 1991 and 1994 on both N-linked carbohydrates and the medullary thyroid carcinoma peptide CGRP shown in Fig. 3(a) and (b) [42].

Whilst most of the Imperial group's activities throughout the 1980s had tended to deviate away from pure sequencing problems, with some notable exceptions [43–46], and into post-translational modification research, many groups worldwide were still attempting to better the performance of triple quads for peptide sequencing by CAD MS/MS, including of course notably Don Hunt's seminal work with John Yates and co-workers [47]. In addition, there were many advocates for high energy CAD on large four-sector instruments [48,49] with the twin arguments that Leu/Ile differentiation was then possible and that more fragmentation led to more certain sequencing. As a group with some of the most extensive sequencing experience worldwide, we at Imperial did not concur with these arguments for two reasons: firstly, we felt that from the biological viewpoint, the Leu/Ile differentiation problem was minor compared to the primary need to determine unknown stretches of protein sequence. In our view, if a sequence proved to be interesting or important, the Leu/Ile assignment could then be determined if necessary in a separate experiment. Secondly, our experience showed that the more simple and more predictable the fragmentation was, then the higher the probability of deducing an unambiguous sequence. High energy fragmentation on four-sector instruments was for us the antithesis of what was required for protein sequencing, and we utilised principally low energy CAD on triple quadrupole instrumentation for our MS/MS analysis, in much the same way as Hunt's laboratory had pioneered. That is not to say that we could not see some advantages in high energy CAD, particularly for carbohydrate studies when we wanted to promote ring fragmentation in order to assign branching structures. For the latter studies we joined the European Large Installation Plan in its application for funding for an extension of the ZAB2SE 2FPD idea, the ZAB-T focal plane MS/MS instrument at the CNR laboratory of Tonino Malorni and Gennaro Marino in Naples. That instrument fulfilled its promise in the carbohydrate area as expected [50,51].

Don Hunt sometimes quipped during our respective Visiting Professorships in Virginia and Imperial that he was only trying to get up to speed in our biopolymer methodology so that one day he would soundly beat us! In 1978 HRM did not expect that to happen any time soon, and to be fair it did take Don 10 years, but at the beginning of the 1990s Don and his group gave us and many others the "coup de grace", not only in technical brilliance but in mirroring our main philosophy of applying mass spectrometry to the most important biochemical and medical problems of the day, with his seminal papers on the identification of MHC Class I and II peptides at high sensitivity using microcapillary-LC–triple quadrupole MS/MS [52,53]. In this work Don's group achieved sensitivities down to low picomole and even subpicomole levels, which although above our then recently reported ZAB FPD data, was on full CAD MS/MS spectra as opposed to just quasi molecular ions with natural fragmentation. This was a significant breakthrough in MS/MS sequencing for which Don and his group have been rightly celebrated, and the immunological field in particular was very excited that Class I and II complex peptide mixtures could now be studied by his new approach. We recognised that the sensitivity enhancement which Don demonstrated in this and follow up papers came from a combination of minimum sample handling coupled with getting the available material to elute in a very narrow (and therefore more intense) peak to maximise the intensity of the MS/MS fragmentation data on the fast scanning Q3. The ionisation method used was of course electrospray, being fully compatible with the liquid eluent from the LC and the low voltage ion source on the first quadrupole. This work was also a wake-up call for magnetic sector specialists such as ourselves, because although the data quality was much better on sector instruments (and we will return to this theme later), that level of sensitivity could not be achieved on four-sector CAD MS/MS instruments, and those instruments were much more expensive than triple quadrupoles. Surprisingly, this message took rather a long time to be appreciated. People continued to try to adapt electrospray sources to high voltage magnetic sector instruments, something which one of us, HRM, had felt was unlikely to be competitive or fully practicable after working on such devices in the mid 1980s during a Visiting Professorship in Life Sciences at Dupont, Wilmington. Don's group went on to drive sensitivity up and sample requirements down even further to the 10 fmol range in subsequent papers, although sequencing officianados including ourselves had in the meantime been

analysing the weaknesses of the triple quadrupole methodology and looking at alternative new instrumentation solutions.

During this period, 1988–1995, one of us (HRM) was senior consultant to one of the main manufacturers, with a remit to advise on biomolecular MS and to help develop new mass spectrometric instrumentation for biopolymer sequencing. Early advice had been to build electrospray triple quadrupole instrumentation as a first priority, rather than magnetic sector possibilities or the increasingly popular MALDI TOF, Ion Trap or FTMS options, and Don Hunt's papers had, we felt, vindicated those decisions. However, at Imperial we also clearly recognised, perhaps better than those with less *de novo* sequencing experience, the weakness of trying to assign unambiguous sequences using unresolved and sometimes statistical triple quadrupole data. Many will not appreciate the often poor quality of such data, but Fig. 4 illustrates the point quite clearly. The data in Fig. 4 show a typical triple quadrupole signal from a peptide study, zoomed in on *m/z* 744, as raw data before smoothing. After smoothing of course the signal looks more acceptable in publications, but the point is that this data is typical of that obtained between the late 1980s and early 1990s at the tens of picomole level on triple quadrupoles. Note that the signal is spread, before smoothing, over several mass units due to Q3 having been set for maximum transmission/low resolution. If we imagine lowering the sample consumption level so that the signal becomes truly statistical, then the relatively few ions generated could arrive during a given short scanning time at any point across the several mass unit window. After smoothing there could be no confidence in the signal actually being nominally 744 as opposed to 773, 775 or even 772 or 744. Since it is crucial to have mass accuracy to differentiate between the single Dalton differences of ASP, ASN, LEU/ILE or GLU and GLN/LYS then the data may become seriously ambiguous depending upon the actual sequence under study.

One effect of Don's pioneering MHC work was to stimulate us to find a different solution to what we saw as an ambiguity problem, whilst preserving the aspects of strategy which we fully agreed with, namely electrospray on a low voltage ion source on a quadrupole mass filter with subsequent low energy CAD. In the high energy versus low energy CAD debate we felt that one rather fundamental point had been missed and our thinking was as follows; detecting interpretable signals is all about signal to noise ratios. If you split the amount of available primary ion signal into an unnecessarily (from the interpretation viewpoint) large number of fragments in the collision cell then it follows that the signal to noise ratios of those fragment ion signals will be poorer than they need have been. Low energy CAD with minimal fragmentation is therefore beneficial to the objective of high sensitivity sequencing. Another aspect of our thinking at the time was to go for minimum path length to minimise signal dispersion and optimise sensitivity and large geometry magnetic sector solutions mitigated against this. An electrospray quadrupole front end to a tandem instrument commensurate with low energy CAD seemed to fulfill the criteria which we felt were important. Whilst triple quad tandem MS/MS provides an ideal platform for certain types of work, for example pharmacokinetic quantitation, and peptide MS/MS up to a certain level, the disadvantages of triple quad MS/MS for ultra-high sensitivity sequencing are summarised in Table 1, the main ones being lack of resolution in the fragment ion mass analysis, coupled with scanning detection which leads to the spending of large amounts of time on areas of the spectrum where no signals may be generated. Our wide angle ZAB 2FPD experiments had demonstrated at the practical level in 1990/1991 that the theoretical gain of up to 100 with non-scanning MCP detection was realisable, but

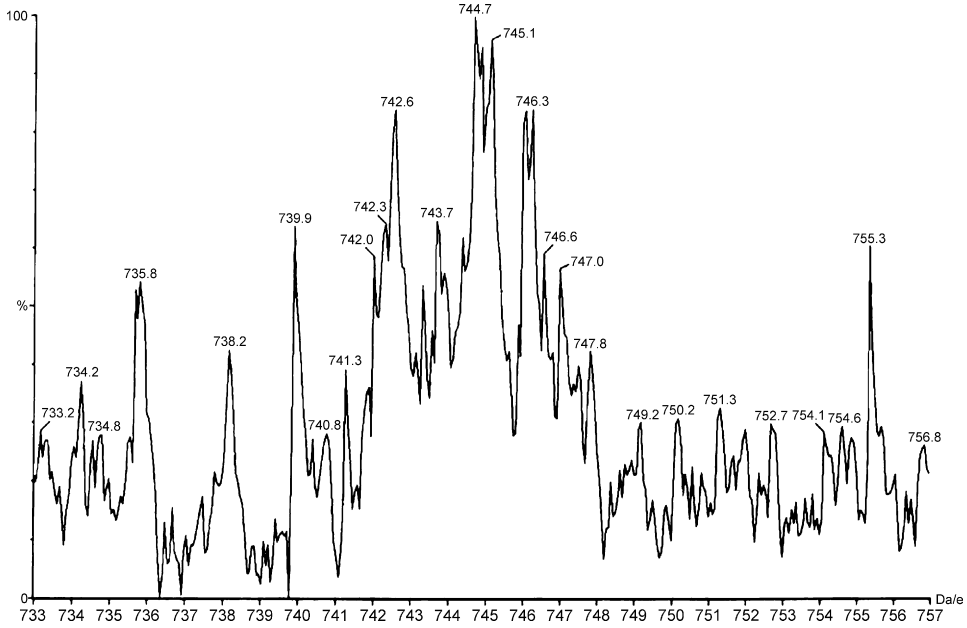

Fig. 4. Close-up of a single signal from a triple quadrupole generated MS/MS spectrum of a peptide in an early 1990s study, seen as raw unsmoothed data, illustrating the wide mass range over which the signal is detected when Q3 has, of necessity, to be set at low resolving power/high transmission. The picture shows the potential (and real) problem of mass assignment if much lower sample loadings into the low femtomole range were to be attempted.

Table 1
Tandem MS/MS triple quadrupole mass spectrometer

Limitations of triple quad geometry for ultra-high sensitivity sequencing (<100 fmol):
Post-CAD scanning detection
Q3 set to low resolution to maximise transmission
Poor mass accuracy when data become statistical
Consequences are limited sensitivity and the possibility of ambiguity in interpretation (e.g., I/L, N, D or K/Q, E).

Table 2
Tandem MS/MS ultra-high sensitivity sequencing initiative

Proposed solution:
A novel geometry quadrupole orthogonal acceleration time of flight Q-TOF instrument
Conceived advantages:
Optimal LC–electrospray-MS1 coupling
Low energy CAD for ease of interpretation
Post-CAD non-scanning detection
Resolved fragment ion spectra with good transmission
Leading to ultra-high sensitivity data with good mass accuracy allowing unambiguous interpretation even of weak MS/MS data.

although some favoured a magnetic sector solution to the additional resolving power problem on any new instrument, we did not, for a combination of the reasons discussed above. The solution to creating the fragment ion resolution and therefore mass accuracy with good transmission lay in a TOF analyser for post CAD analysis and MCP detection, and in particular an orthogonal analyser [54] fitted with a reflectron. This novel Q-TOF combination equipped with electrospray was our favoured solution and our suggestion to resolve the interpretation ambiguity problem which we saw in triple quadrupole data, Table 2. Initially the manufacturers favoured the building of a magnetic sector o-TOF option, and it took another couple of years of thinking to finally convince ourselves and others that the Q-TOF should definitely be built. Finally, in 1994 persuasion won the day, and a prototype instrument was developed. The results from that pre-production instrument were, as we expected, spectacular [55], giving subfemtomole sensitivity, but crucially, together with resolution and thus mass accuracy as shown in some of that initial research in Fig. 5.

The first application of the Q-TOF to the field of glycoproteomics was a collaboration with Chris West's group in Florida on an interesting potentially *cytoplasmic* glycosylation of a protein named FP21, found in the ubiquitinylation F-box involved in cell cycle regulation. West's group had found evidence of radiolabeled fucose incorporation into the molecule, suggesting the presence of glycosylation. Since the protein sequence was known, our strategy was to calculate expected Lys-C digest peptide masses, adding sugar masses onto the doubly or triply charged ions. Candidate glycopeptides would then be MS/MSed to define structure. This approach drew a blank, although it was noticed that one of the expected peptides was absent from the mapping experiment (it is not unusual for a mapping experiment not to cover 100% of a protein molecule). However, the Q-TOF data showed a strong candidate signal at *m/z* 829.35 (3+) which did not map onto the known FP21 structure, and in addition the triply charged signal had higher mass doubly charged satellites apparently corresponding to sugar additions as seen in Fig. 6. Disappointingly, the subtraction of these signals did not leave a mass corresponding to any of the Lys-C peptide masses in FP21! The conundrum was solved by careful CAD MS/MS analysis of the triply charged 829.35 ion on the Q-TOF, generating the

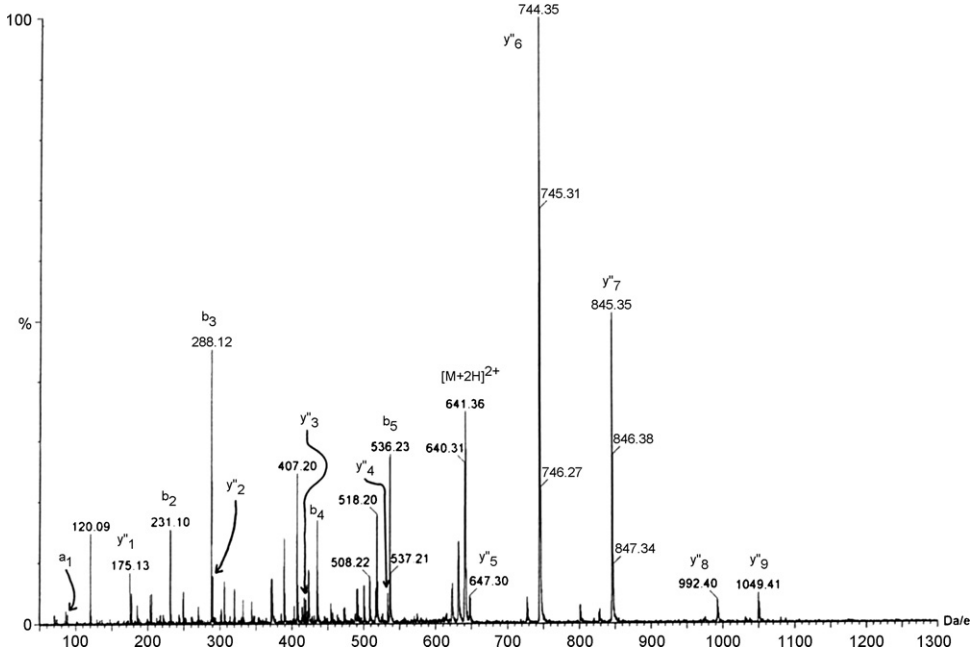

Fig. 5. Original research data from the prototype Q-TOF instrument [55] showing that isotopically resolved MS/MS data with good mass accuracy was now achievable using this novel geometry. The spectrum shows the MS/MS of *m/z* 640.2^{2+} derived from an in-gel tryptic digest. The sequence assigned as DDGFTPNNEDR identified the 19 kDa band as Gene B protein.

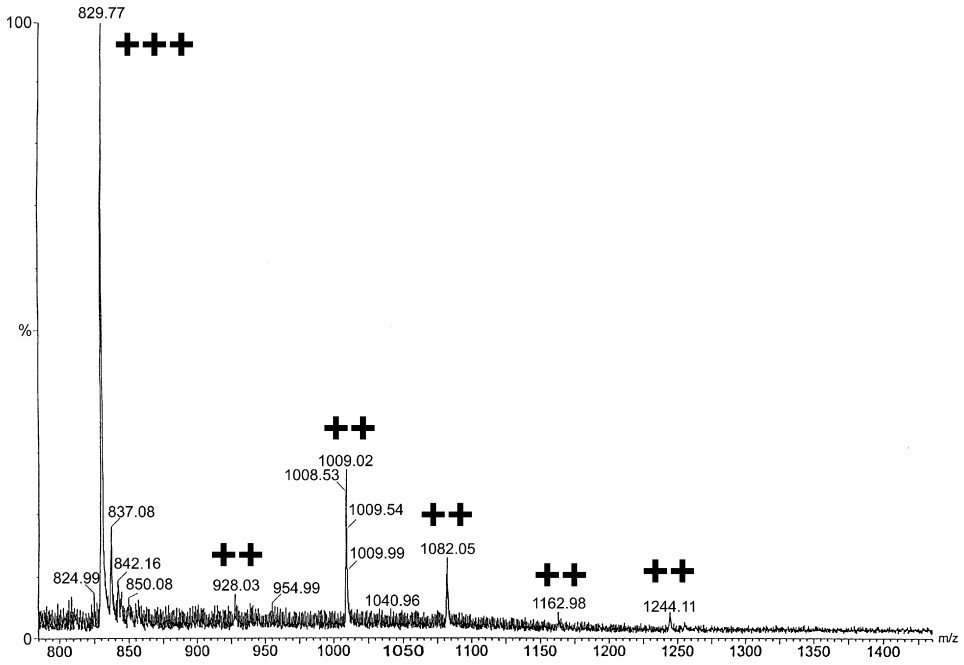

Fig. 6. The MS spectrum of an LC fraction from a Lys C digest of FP21 (Skp 1) protein from *D. discoidium* taken on the Q-TOF [56], showing the discovery of glycosylation in the molecule from the mass differences seen between low level doubly charged fragment ions created by cone-induced fragmentation.

spectrum shown in Fig. 7. These data immediately allowed assignment of the peptide incorporated in the signal as residues 139 to 151 of the FP21 protein via the b and y ions present at *m/z* 230, 303, 377, 416, 478, 544, etc. However, the y ions beyond *m/z* 1060 were 16 Da higher than predicted. Our interpretation was that a double post-translational modification of this protein had taken place, leading first to the creation of a hydroxyproline at position 143 in the sequence, then modified by a novel O-linked pentasaccharide glycosylation containing a linking fucose at this position [56]. Note not only the clarity of the well resolved signals in Fig. 7, but also the important ions at *m/z* 1173.64 and 1376.67 (a HexNAc apart) which prove that the sugar attachment is at the hydroxyproline position.

2. Current and future research

The powerful combination of the specific cleavage/mass mapping strategies are further enhanced in on-line nano-LC–MS

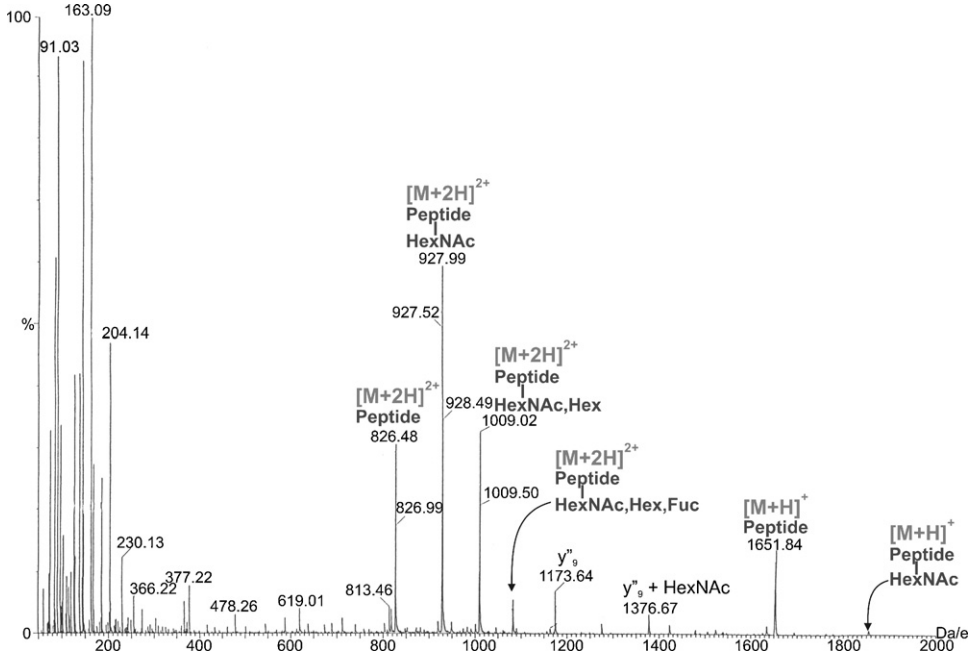

Fig. 7. The partial Q-TOF MS/MS spectrum of the 829.35^{3+} ion in Fig. 6 showing the 1173 fragment ion assigned as a hydroxyproline difference from *m/z* 1060, and the signal at *m/z* 1376 providing evidence of the carbohydrate attachment at hydroxyproline in position 5 of the assigned sequence NDFTP(OH)EEEEQIRK.

and MS/MS experiments on the Q-TOF, and have been applied in our laboratory to numerous challenging O-linked analyses, including characterising the multiple site-specific glycosylation of the MPB83 tuberculosis antigen of *Mycobacterium bovis*, found to contain Man and Man alpha 1–3 Man on adjacent threonines [57], and defining the changes in O-linked glycosylation of CD8 beta associated with T cell maturation in the thymus [58].

Much of the group's work is on the much larger structures associated with N-linked glycosylation, where as mentioned earlier there are text book biosynthetic pathways to assist in assigning possible structures to mass spectrometric data. Principal amongst these rules in eukaryotes (which until recently were thought to be the only organisms with N-linked structures) is the sacrosanct "trimannosyl core" attachment of the N-linked oligosaccharide to the asparagine residue in the protein via two GlcNAc residues. When our collaborators in ETH Zurich and the London School of Hygiene and Tropical Medicine contacted us with gene-derived evidence of the presence of *N*-glycosylation machinery in *Campylobacter jejuni*, we were initially sceptical that the predicted consensus sequences would be occupied. However, the advantage of keeping an open mind in structural biology can never be over-emphasised to new students! A nano-LC–MS mapping experiment on a in-gel tryptic digest of protein Peb3 from *C. jejuni* left a number of signals unassigned, most of which turned out to be chymotryptic splits. However, a relatively low mass doubly charged signal at *m/z* 1057.5 gave a very interesting MS/MS spectrum on the Q-TOF as seen in Fig. 8. The fragmentation looked to be classical O-linked rather than N-linked, but peptide fragment ions defined the peptide as DFNVSK, containing a eukaryotic type consensus sequence NVS, and they also showed that the sugar linkage was on the asparagine rather than the serine. Even more interestingly, the interpretation of the fragmentation pattern defined not only a novel sugar linkage 2,4-diacetamido-2,4,6-trideoxyhexose to the Asn residue, but also a very different overall heptasaccharide oligosaccharide structure compared to that seen in eukaryotes (Fig. 8) [59]. This discovery of a new type of *N*-glycosylation has recently been extended by using the mass spectrometric screening strategy to carry out functional genomic studies [60] and in bio-engineering N-linked proteins carrying O antigen lipopolysaccharide structures in *Escherichia coli* [61].

A particular current research interest of the group is the mechanism of fertilisation. Mammalian sperm initiate fertilisation by binding to the specialized extracellular matrix of eggs known as the zona pellucida (ZP). Murine ZP consists of three major glycoproteins, mZP1, mZP2 and mZP3. Glycans on the latter are believed to be responsible for initial sperm–egg binding and the induction of the acrosome reaction. Over the past decade a variety of genetic, biophysical, and biochemical techniques have been employed to investigate the putative sperm binding glycans but their structures remain elusive. Sequences have been established for the majority of *N*- and *O*-glycans on mZP3 [62–64] and sites of glycosylation are also known [65]. The challenge now is to determine which glycans are located at each of the glycosylation sites and to establish which glycoforms are involved in sperm binding. Previously we have employed glycoproteomic approaches to identify the *O*-glycans attached at two sites in a conserved domain of ZP3 [66]. Here we present data on the glycan repertoire at one of the *N*-glycosylation sites. We describe current strategies for manual interpretation of complex glycoproteomic data and report on the progress towards automated interpretation of these data.

mZP3 was digested with trypsin and subjected to online-nano-LC–ES–MS and data-dependent MS/MS. Diagnostic sugar fragment ions in the MS data indicated that glycopeptides

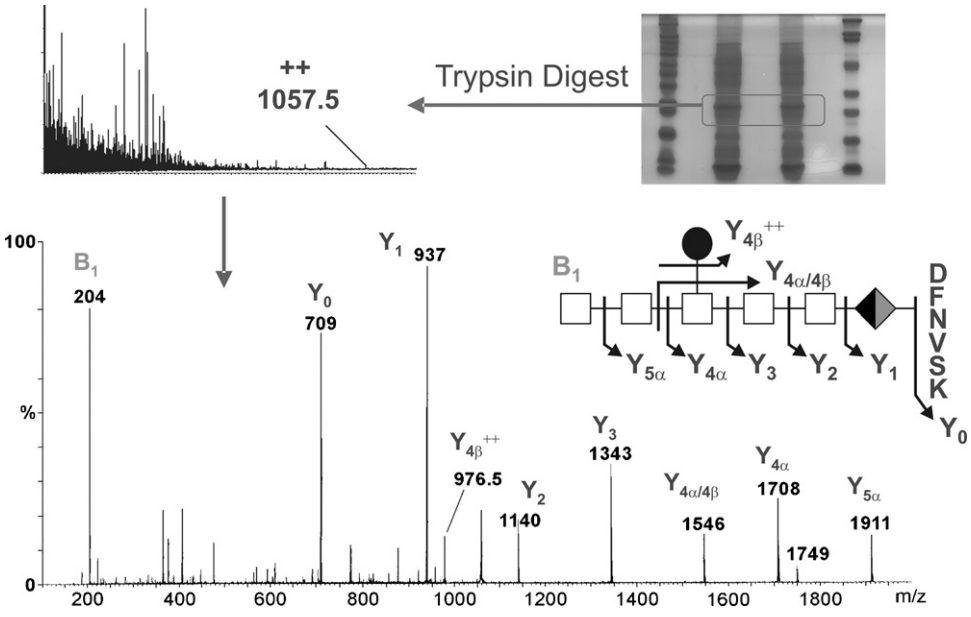

Fig. 8. The gel band, MS spectrum and Q-Star MS/MS spectrum of the *m/z* 1057^{2+} signal from an in-gel tryptic digest, showing the discovery of a new type of N-linked glycosylation in a study of proteins from the pathological organism *C. jejuni* [59]. Symbolic nomenclature used is that of the Consortium for Functional Glycomics (CFG, www.functionalglycomics.org): white square is GalNAc, black circle is Glc, split-diamond is 2,4-diacetamido-2,4,6-trideoxyhexose (DATDH).

were eluting around 60 min. MS data from 58.7 to 59.2 min were summed for detailed examination (Fig. 9). Abundant fragment ions corresponding to sugar sequences were observed at *m/z* 366 (HexHexNAc), *m/z* 407 ($HexNAc_2$), *m/z* 528 ($Hex_2HexNAc$), *m/z* 657 (NeuAcHexHexNAc), *m/z* 673 (NeuGcHexHexNAc), *m/z* 731 ($Hex_2HexNAc_2$), *m/z* 860 ($NeuAcHexHexNAc_2$) and *m/z* 876 ($NeuGcHexHexNAc_2$). Based on earlier work from our laboratory [64], the compositions $Hex_2HexNAc^+$ and $NeuAc/NeuGcHexHexNAc_2^+$ most likely correspond to the Galα1-3Galβ1-4GlcNAc sequence and the Sd^a determinant (GalNAcβ1-4[NeuAc/NeuGcα2-3]Galβ1-4GlcNAc, respectively.

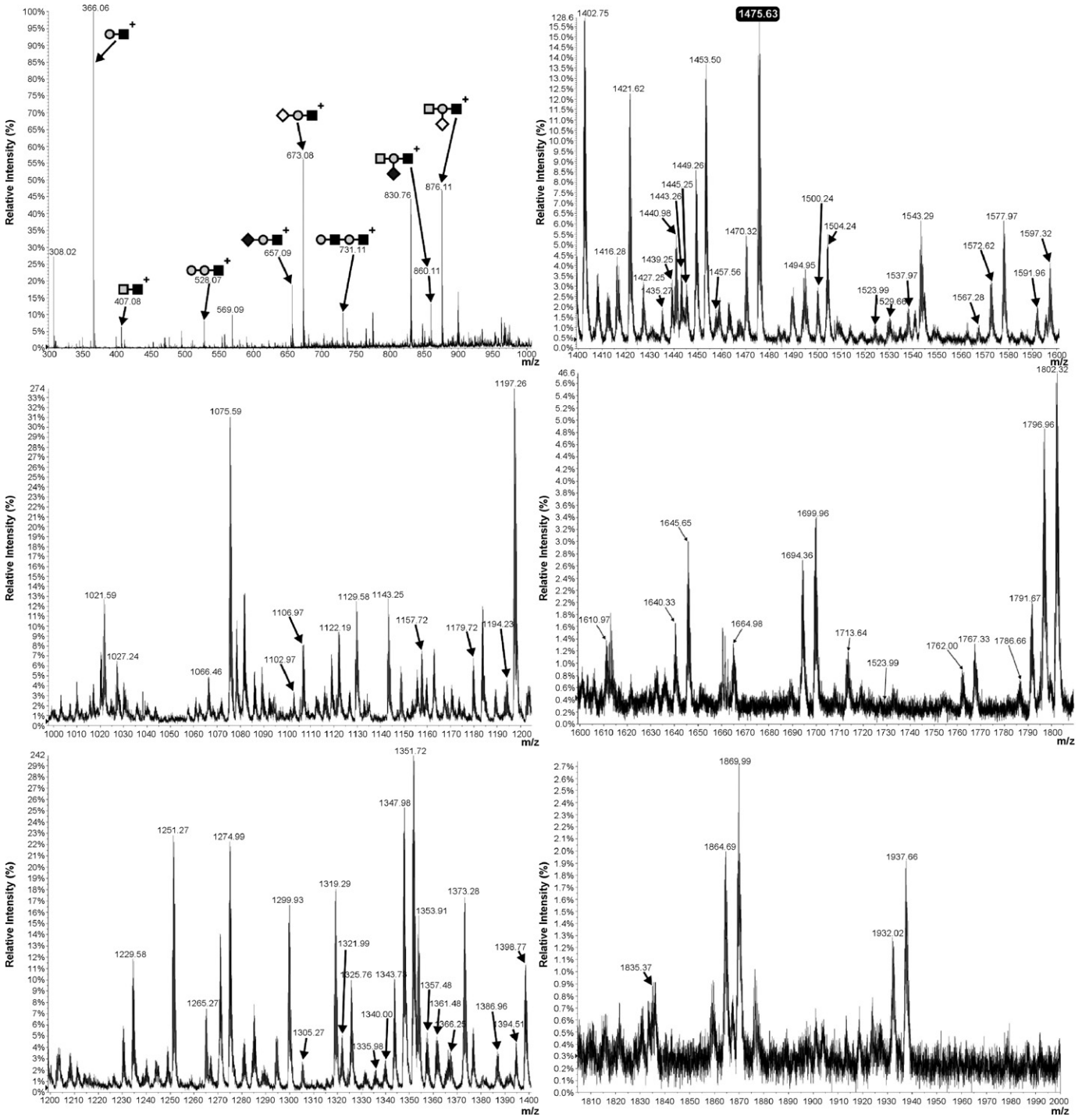

Fig. 9. Summed MS data acquired between 58.7 and 59.2 min in the nano-LC–ES–MS analysis of a tryptic digest of ZP3. The six panels represent sequential mass ranges of the same spectrum, chosen to best highlight the glycan and glycopeptide peaks. Sugar fragment ions which were formed during ionisation are observed in the low mass portion of the spectrum and their compositions are shown in the cartoon annotations using symbols employed by the CFG: white circle is Gal, black square is GlcNAc, white square is GalNAc, black diamond is NeuAc, white diamond is NeuGc. Molecular ions attributable to glycopeptides are annotated with their *m/z* values. Their charges are given in Table 3.

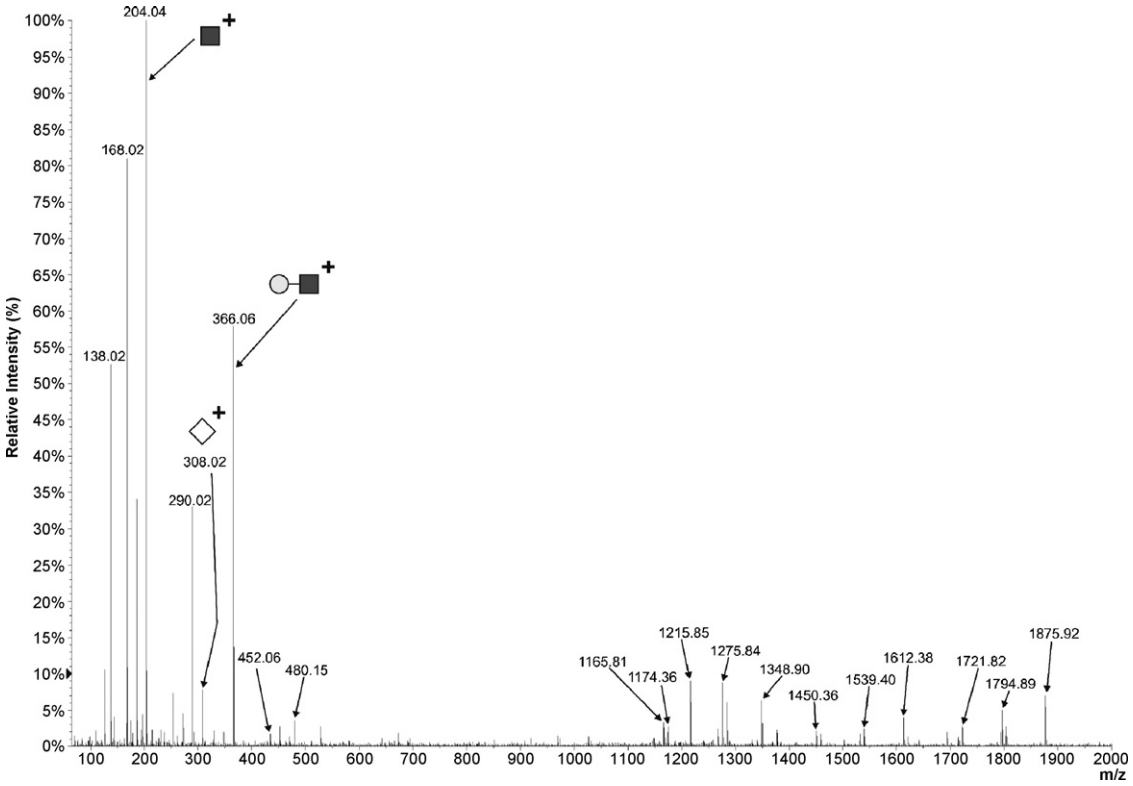

Fig. 10. The triply charged molecular ion at *m/z* 1475.63 (highlighted in panel 4 of Fig. 9) was selected by the data-dependent software for collisional activation and the MS/MS spectrum is shown in this figure. The GlcNAc and NeuGc fragment ions at *m/z* 204 and 308, respectively, are accompanied by secondary fragment ions corresponding to water losses and ring cleavages. The annotated ions above *m/z* 1100 are all doubly charged (for assignments see the text).

The presence of both NeuAc and NeuGc in the sialylated glycans was exploited whilst searching for glycoforms. Thus the mass spectrum (Fig. 9) was searched for pairs of doubly, triply and quadruply charged ions differing by *m/z* 8, 5.3 and 4, respectively. For example the pair of triply charged ions at *m/z* 1572.62 (M_r 4714.9) and *m/z* 1577.97 (M_r 4730.9) that are 5.32 amu apart, indicates that the latter structure contains at least one NeuGc residue whereas the former contains at least one NeuAc.

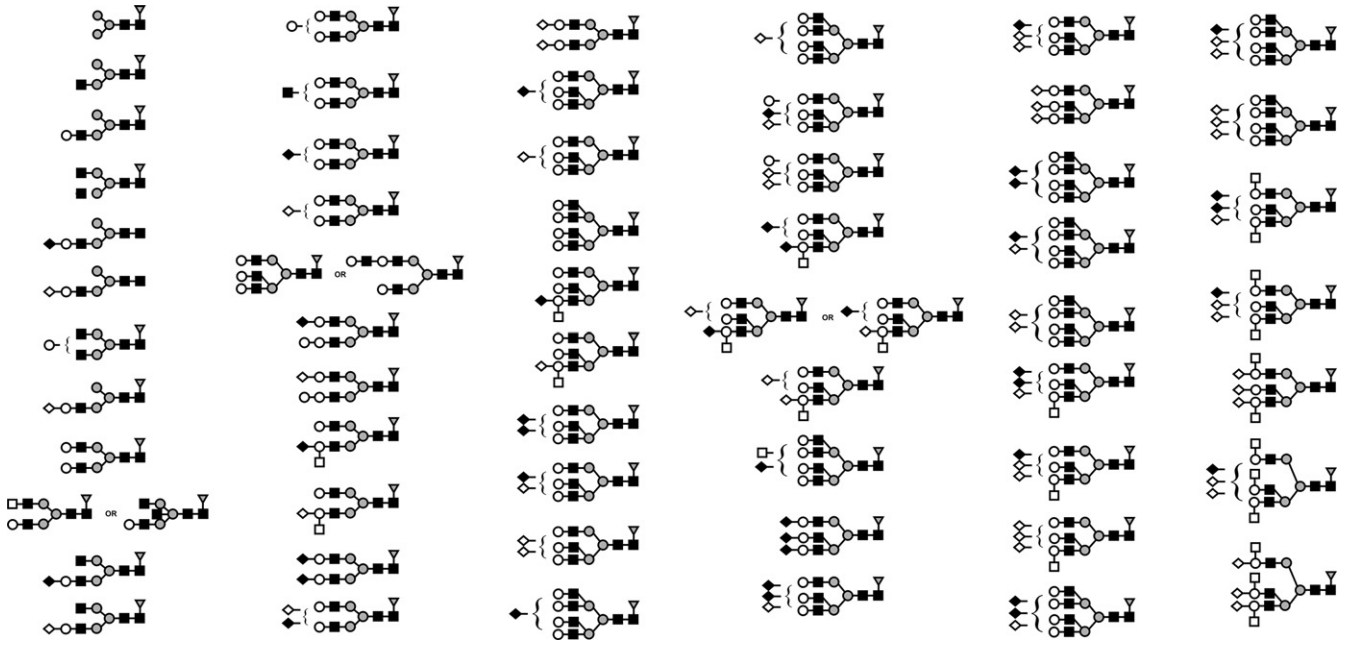

Fig. 11. Structures of the glycans whose compositions are given in Table 3. These were deduced taking into account prior knowledge of antennae sequences in ZP3 [64]. Symbolic nomenclature used is that of the CFG: grey triangle is Fuc, white square is GalNAc, white circle is Gal, black square is GlcNAc, black circle is Glc, black diamond is NeuAc, white diamond is NeuGc.

In order to deduce the compositions of these carbohydrate structures, it was necessary to extrapolate from the data by either adding or subtracting various sugar mass residues to the starting structure and then searching for the multiply charged ions of these glycans in the mass spectrum. For instance, by subtracting the mass of a NeuGc residue from the total molecular mass of the NeuGc-containing glycopeptide at M_r 4730.9, a structure with an expected molecular mass of 4423.8 should be detected in the spectrum. It is indeed observed as a triply charged species at 1475.63. Fortunately this ion had been selected for data-dependent CAD at 59.2 min and the resulting MS/MS spectrum (Fig. 10) provided definitive evidence for glycosylation via the presence of *m/z* 204 ($HexNAc^+$), 308 ($NeuGc^+$), 366 ($HexHexNAc^+$) and 673 ($NeuGcHexHexNAc^+$). The presence of sialylated fragment ions indicates that at least one antenna carries NeuGc. This therefore signifies that the component at 1577.97^{3+} must contain at least two NeuGc residues. The presence of a molecular ion at *m/z* 1373.28^{3+} (Fig. 9), which corresponds to a glycopeptide with two fewer NeuGc residues than *m/z* 1577.97^{3+}, provided further evidence for this conclusion. Together with *m/z* 1475.63^{3+}, *m/z* 1373.28^{3+} had also been selected for CAD MS/MS giving fragment ions at *m/z* 204 and 366 for HexNAc and HexHexNAc, respectively (data not shown). Since no signals for sialylated fragments were observed in this MS/MS spectrum, it can be assumed that the glycopeptide of M_r 4730.9 carries two NeuGc residues. By repeating this process through the addition and subtraction of other sugar masses, it was deduced that the most likely composition of the glycoform at 1577.97^{3+} is a bi-antennary, core-fucosylated glycan carrying two NeuGc residues, i.e., $NeuGc_2FucHex_5HexNAc_4$. Subtracting the mass for this carbohydrate structure (2382.8 U) from the calculated molecular weight of the 1577.97^{3+} ion (M_r 4730.9) gives a residual mass of 2348.1 Da for the peptide component of the glycopeptide. This peptide maps to the mZP3 sequence $^{(257)}$PRPETLQFTVDVFHFANSSR$^{(276)}$ which has a consensus glycosylation site at Asn273. Evidence that the glycosylated peptide had been correctly identified was provided by the clusters of doubly charged fragment ions observed between *m/z* 1100 and 1900 in the MS/MS spectrum of *m/z* 1475.63 (Fig. 10). Thus *m/z* 1174.36^{2+} is the predicted mass for the intact peptide after loss of the glycan, whilst the signals from *m/z* 1275.84^{2+} to 1875.92^{2+} are increments of HexNAc, Fuc or Hex higher. The signal at *m/z* 1215.85^{2+} is assigned to a fragment ion arising from cross ring cleavage of the Asn-linked HexNAc residue, whilst *m/z* 1165.81^{2+} is consistent with cleavage on the acyl side of the *N*-glycosidic bond during glycan loss. Moreover, the *N*-terminal sequence of the peptide was confirmed by the b-ions at *m/z* 254 (PR) and 480 (PRPE), with the latter accompanied by an a-ion at *m/z* 452.

By applying this data analysis strategy to other multiply charged signals in the spectrum shown in Fig. 9, a total of fifty eight glycan compositions were assigned (Table 3) to the Asn273 glycoforms. Likely structures for these glycans, based on their compositions and previous glycomic analysis of ZP3 [64], are shown in Fig. 11.

We are developing prototype software for automatically analyzing glycopeptide spectra, which we applied to the same

Table 3
This table gives the glycan compositions of the PRPETLQFTVDVFHFANSSR glycopeptides which were manually assigned to triply or quadruply charged molecular ions observed in the spectrum reproduced in Fig. 9

M_r	*m/z*	a	b	c	d	e
3385.8	1129.6^{3+}	0	0	1	3	2
3588.8	1197.3^{3+}	0	0	1	3	3
3750.8	1251.3^{3+}	0	0	1	4	3
3792.8	1265.3^{3+}	0	0	1	3	4
3896.8	1299.9^{3+}	0	1	1	3	3
3912.8	1305.3^{3+}	0	0	1	5	3
3954.8	1319.3^{3+}	0	0	1	4	4
4058.8	1353.9^{3+}	0	1	1	4	3
4116.9	1373.3^{3+}	0	0	1	5	4
4157.9	1387.0^{3+}	0	0	1	4	5
4245.8	1416.3^{3+}	1	0	1	4	4
4261.9	1421.6^{3+}	0	1	1	4	4
4278.8	1427.3^{3+}	0	0	1	6	4
4319.9	1441.0^{3+}	0	0	1	5	5
4407.9	1470.3^{3+}	1	0	1	5	4
4423.9	1475.6^{3+}	0	1	1	5	4
4481.9	1495.0^{3+}	0	0	1	6	5
4567.8	1523.6^{3+}	1	0	1	6	4
4584.9	1529.3^{3+}	0	1	1	6	4
4619.0	1538.0^{3+}	1	0	1	5	5
4646.9	1543.3^{3+}	0	1	1	5	5
4698.9	1567.3^{3+}	2	0	1	5	4
4714.9	1572.6^{3+}	1	1	1	5	4
4730.9	1578.0^{3+}	0	2	1	5	4
4772.9	1592.0^{3+}	1	0	1	6	5
4788.9	1597.3^{3+}	0	1	1	6	5
4845.9	1616.3^{3+}	0	0	1	7	6
4974.9	1659.3^{3+}	1	0	1	6	6
4992.0	1665.0^{3+}	0	1	1	6	6
5065.0	1689.3^{3+}	2	0	1	6	5
5080.0	1694.3^{3+}	1	1	1	6	5
5096.9	1700.0^{3+}	0	2	1	6	5
5137.9	1713.6^{3+}	1	0	1	7	6
5153.9	1719.0^{3+}	0	1	1	7	6
5241.1	1311.3^{4+}	1	1	1	7	5
5258.0	1315.5^{4+}	0	2	1	7	5
5267.8	1318.0^{4+}	2	0	1	6	6
5284.0	1762.0^{3+}	1	1	1	6	6
5299.0	1767.4^{3+}	0	2	1	6	6
5339.0	1335.7^{4+}	1	0	1	7	7
5355.0	1786.0^{3+}	3	0	1	6	5
5371.0	1791.7^{3+}	2	1	1	6	5
5387.9	1797.0^{3+}	1	2	1	6	5
5403.9	1802.3^{3+}	0	3	1	6	5
5425.1	1357.3^{4+}	2	0	1	7	6
5443.9	1362.0^{4+}	1	1	1	7	6
5461.0	1366.2^{4+}	0	2	1	7	6
5575.0	1394.8^{4+}	2	1	1	6	6
5591.0	1398.8^{4+}	1	2	1	6	6
5607.0	1402.7^{4+}	0	3	1	6	6
5737.1	1435.3^{4+}	2	1	1	7	6
5753.0	1439.3^{4+}	1	2	1	7	6
5769.0	1443.3^{4+}	0	3	1	7	6
5777.1	1445.3^{4+}	2	1	1	6	7
5794.0	1449.5^{4+}	1	2	1	6	7
5810.0	1453.5^{4+}	0	3	1	6	7
5997.0	1500.3^{4+}	1	2	1	6	8
6013.0	1504.3^{4+}	0	3	1	6	8

The *m/z* values in the table correspond to the most abundant isotope in each cluster and the M_r value is calculated accordingly. Columns a, b, c, d, and e are NeuAc, NeuGc, Fuc, Hex and HexNAc, respectively. Assignments are made after choosing ^{12}C masses, calibration corrected using regional peptide assignments, and cross-checked for acceptable isotopic distribution.

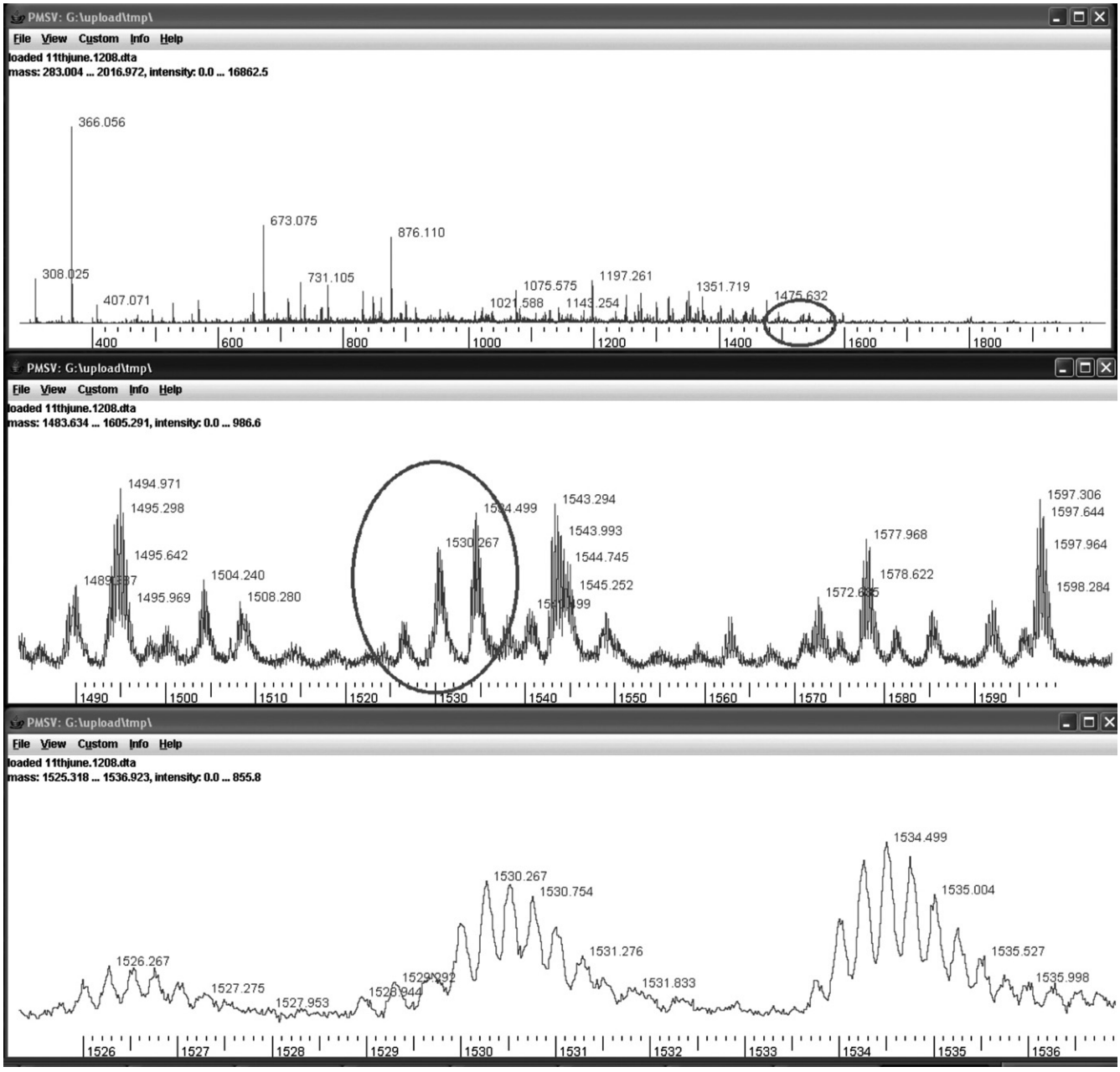

Fig. 12. Three views of MS data summed between 57.5 and 59.5 min. The bottom two panels show increasingly zoomed in views of the peak at 1534, which corresponds to mZP3:257–276 + $NeuGc_3FucHex_8HexNAc_7$. The peaks corresponding to the $NeuAc_2NeuGc$ and $NeuAcNeuGc_2$ analogues are also visible.

spectra whose manual annotation is described above. The software examines each spectrum in an LC/MS series, hunting for isotope envelopes of ions that have sufficient resolution to determine the charge, and that have the appropriate shape for a glycopeptide. For each envelope found, the program then determines which peak is the +0 isotope, and converts it from *m/z* to a mass. Mimicking the strategy of human annotators might involve looking for pairs of masses that differ by the mass of a glycan. But the software takes a different approach. It builds a table containing the masses of all tryptic glycopeptides from the target molecules (in this case mZP2 and mZP3) and then for each mass computed from the LC/MS spectra checks for a matching target mass in the table.

One of the major advantages of automatic software is that it can easily find glycopeptides that are missed in the tedious process of human annotation (Fig. 12 gives an example). The envelope at *m/z* 1534 was examined by the program, and as shown in Fig. 13 has a shape very similar to the predicted envelope of a "generic" glycopeptide ion of charge +4. After verifying the similarity, the program then found the +0 isotope peak at 1533.76 and computed the total mass as 6131.0 which matched (after correcting for calibration) the target glycopeptide mZP3:257–276 + $NeuGc_3FucHex_8HexNAc_7$ (where mZP3:257–276 is the sequence PRPETLQFTVDVFHFANSSR) of mass 6132.48. Although this particular glycoform was not identified by manual annotation, the spec-

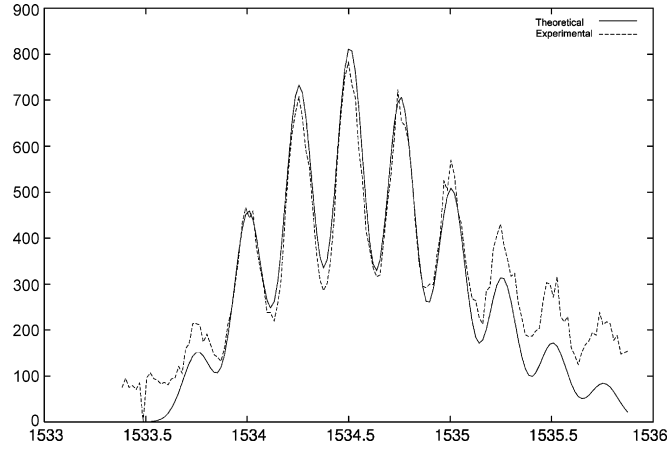

Fig. 13. The solid line shows the theoretical spectrum of a quadruply charged glycopeptide of mass $m = 6132.48$. This is formed by summing Gaussians centered at $(m+4)/4$, $(m+5)/4$, $(m+6)/4$, etc., with each Gaussian weighted by the theoretical abundance of the corresponding isotopic species. The dashed line is the experimental data, summed over the spectra between 57.5 and 59.5 min.

trum also contains peaks corresponding to $NeuAc_2NeuGc$ and $NeuAcNeuGc_2$, which is strong confirmation for this glycopeptide. Altogether the current version of the program found 27 additional glycopeptides not present in Table 3 which had either low mass or that had at least one NeuAc/NeuGc analogue. These

Table 4
This table gives the glycan compositions of the PRPETLQFTVDVFHFANSSR glycopeptides which were automatically found by the prototype software

M_r	m/z	a	b	c	d	e
3223.49	1075.260^{3+}	0	0	1	2	2
3426.57	1142.930^{3+}	0	0	1	2	3
3442.57	1148.260^{3+}	0	0	0	3	3
3912.73	1304.950^{3+}	0	0	1	5	3
4933.07	1233.98^{4+}	0	2	1	5	5
4917.07	1229.990^{4+}	1	1	1	5	5
6074.43	1519.250^{4+}	0	4	1	7	6
6058.44	1515.250^{4+}	1	3	1	7	6
6042.44	1511.250^{4+}	2	2	1	7	6
5663.33	1416.500^{4+}	0	2	1	7	7
5647.34	1412.510^{4+}	1	1	1	7	7
6132.48	1533.760^{4+}	0	3	1	8	7
6100.49	1525.760^{4+}	2	1	1	8	7
6116.48	1529.760^{4+}	1	2	1	8	7
6439.57	1610.510^{4+}	0	4	1	8	7
6423.57	1606.510^{4+}	1	3	1	8	7
6407.58	1602.520^{4+}	2	2	1	8	7
6480.59	1620.750^{4+}	0	4	1	7	8
6464.60	1616.760^{4+}	1	3	1	7	8
6335.55	1584.510^{4+}	0	3	1	8	8
6319.56	1580.520^{4+}	1	2	1	8	8
6190.52	1548.260^{4+}	0	2	1	9	8
6174.52	1544.270^{4+}	1	1	1	9	8
6481.61	1621.020^{4+}	1	2	1	9	8
6497.61	1625.020^{4+}	0	3	1	9	8
7610.06	1522.710^{5+}	3	1	1	8	13
7626.05	1271.760^{6+}	2	2	1	8	13

The m/z values in the table correspond to the estimated +0 (^{12}C) isotope in each cluster and the M_r value here is the theoretical mass of the glycopeptide. Columns a, b, c, d, and e are NeuAc, NeuGc, Fuc, Hex and HexNAc, respectively.

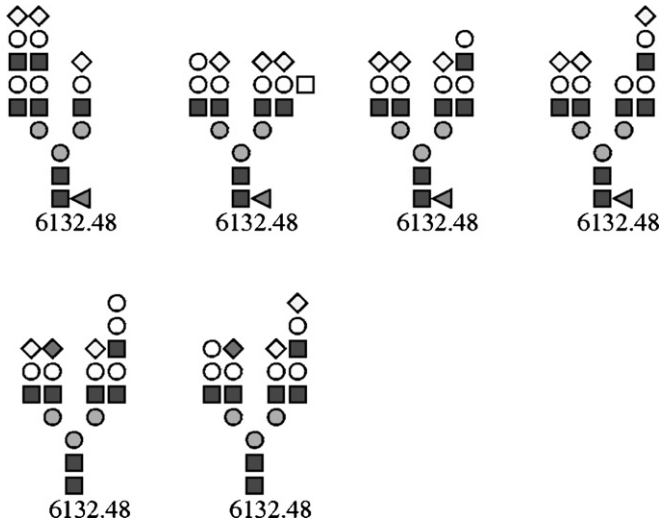

Fig. 14. The most plausible cartoons for the glycans of the glycopeptide of mass 6132.48. The first row shows the highest ranked cartoons for $NeuGc_3FucHex_8HexNAc_7$, and the second row shows the highest for the less likely composition $NeuAcNeuGc_2Hex_9HexNAc_7$. Symbolic nomenclature used is that of the CFG: grey triangle is Fuc, white square is GalNAc, white circle is Gal, black square is GlcNAc, black diamond is NeuAc, white diamond is NeuGc.

are shown in Table 4. We expect improved versions to find additional glycoforms.

Once a peak has been identified as having the mass of a tryptic glycopeptide, the next step is to assign a plausible cartoon to the glycan. This is done using an improved version of the Cartoonist algorithm [68]. The algorithm has two parts: biosynthetic rules and demerits. The biosynthetic rules are used to automatically generate about 145,000 cartoons, which should be a superset of all biologically possible glycans. Demerits are used to rank cartoons of the same composition by downgrading the less likely ones, and can be adjusted for different animals and tissues. The demerits used for the mouse ZP3 glycans take into account prior glycomics work [64] and downgrade structures such as bisecting GlcNAc, LacdiNAc antennae and peripheral fucosylation. The peak at 1534 (corresponding to total mass 6132.48) has two possible glycan compositions: $NeuGc_3FucHex_8HexNAc_7$ and $NeuAcNeuGc_2Hex_9HexNAc_7$. The highest ranked cartoons for each composition are shown in Fig. 14.

This software, whilst not yet fully developed, promises to supersede current manual interpretation of complex glycopeptide data sets or at least to augment it, although as we have illustrated above with the *D. discoidium* FP21 and *C. jejuni* PEB studies, biomolecular post-translational research still throws up problems which are likely to need the intervention of human intelligence!

Acknowledgements

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) and the Wellcome Trust (to A.D. and H.R.M.) and grants from the National Institutes of Health (HD 35652 to G.F.C. via an institutional incentive fund) and the Breeden-Adams Cancer Foundation (G.F.C.). A.D.

is a BBSRC Professorial Fellow, and S.C. was supported by a BBSRC Studentship. D.G. was partially supported by NIH Grant R01GM074128 from the NIGMS. Grateful thanks are due to Dr. Simon North for his efforts in reproducing some of the historical figures.

Appendix A

A.1. Materials and methods

A.1.1. Materials

Purified ZP was isolated from flash-frozen mouse ovaries obtained from 12- to 14-week-old wild-type mice and ZP3 was separated from ZP1 and ZP2 by gel filtration as described previously [67]. All other chemicals and reagents were purchased from Sigma unless otherwise indicated.

A.1.2. Nano-LC–ES–MS/MS analysis

Tryptic digests were analyzed by nano-LC–ES–MS/MS using a reverse-phase nano-HPLC system (Dionex, Sunnyvale, CA, USA) connected to a quadrupole TOF mass spectrometer (Q-STAR Pulsar I, MDS Sciex). The digests were separated by a binary nano-HPLC gradient generated by an ultimate pump fitted with a Famos autosampler and a Switchos microcolumn switching module (LC Packings, Amsterdam, The Netherlands). An analytical C18 nanocapillary (75 mm inside diameter × 15 cm, PepMap) and a micro precolumn C18 cartridge were employed for on-line peptide separation. The digest was first loaded onto the precolumn and eluted with 0.1% formic acid (Sigma) in water (HPLC grade, Purite) for 4 min. The eluent was then transferred onto an analytical C18 nanocapillary HPLC column and eluted at a flow rate of 150 nL/min using the following gradient of solvent A [0.05%, v/v formic acid in a 95:5, v/v water/acetonitrile mixture] and solvent B [0.04% formic acid in a 95:5, v/v acetonitrile/water mixture]: 99% A from 0 to 5 min, 99–90% A from 5 to 10 min, 90–60% A from 10 to 70 min, 60–50% A from 70 to 71 min, 50–5% A from 71 to 75 min, 5% A from 75 to 85 min, 5–95% A from 85 to 86 min, and 95% A from 86 to 90 min. Data acquisition was performed using Analyst QS software with an automatic information-dependent-acquisition (IDA) function.

A.1.3. Algorithms

Isotope envelopes are found by fitting a sum of Gaussians to the spectrum, using the nonlinear minimisation routine dfpmin from numerical recipes [69]. The optimization is over three scalar parameters: the position of the series of Gaussians, the width of the Gaussians, and an overall scale factor. The Gaussians all have the same width (variance), and a single scale factor determines all their heights because their relative heights are precomputed based on isotope abundances. These are computed using the total mass corresponding to the peak, and assuming a generic glycopeptide consisting of 50% H atoms, 30% Carbon, 5% Nitrogen and 15% Oxygen.

The spectrum is centroided, and the resulting peaks are considered one at a time starting with the most intense. The charge of a peak is estimated by looking at nearby peaks, and the resulting region of the spectrum is fitted to a sum of Gaussians spaced 1/charge apart. The starting point for dfpmin iteration sets the position parameter to match the Gaussian of highest theoretical abundance to the highest observed peak in the region. Then dfpmin is called two additional times, shifted 1/charge to the left, and to the right.

The table of target glycopeptide masses is computed using Cartoonist [68]. Cartoonist generates about 145,000 cartoons using biosynthetic rules, which build up each antenna as a series of lactosamine units followed by one of nine capping units. Corresponding to these cartoons are approximately 6000 different glycan compositions. The masses of these compositions are used to populate the table of target masses.

To determine the tolerance used in matching the location (mass) of the +0 Gaussian and the table of targeted masses, the entire set of spectra are first scanned for unmodified peptides that are validated by an MS/MS spectrum. The difference between the observed and theoretical masses of these peptides is used to recalibrate the spectrum as described in [70], and the deviation after recalibration is used as the tolerance. For the spectra used here, that tolerance is 0.05 Da.

For conversion to cartoons, the demerits system of Cartoonist (see main text) was used [68,70]. Demerits were applied to cartoons containing anything from the following list: Sialyl Lewis X, LacdiNAc, sialyl LacdiNAc, two stacked sialic acids, an antennae with exactly 2 mannose, bisecting GlcNAc, an antenna with multiple fucose, an absence of base fucosylation, five antennae, a hybrid structure, uncapped GlcNac, a single long antennae, or lack of a full trimannosyl core.

References

[1] H.R. Morris, A.J. Geddes, G.N. Graham, Biochem. J. 111 (1969) 38.
[2] A.J. Geddes, G.N. Graham, H.R. Morris, F. Lucas, M. Barber, W.A. Wolstenholme, Biochem. J. 114 (1969) 695.
[3] H.R. Morris, D.H. Williams, R.P. Ambler, Biochem. J. 125 (1971) 189.
[4] H.R. Morris, D.H. Williams, J. Chem. Soc., Chem. Commun. (1972) 114.
[5] A. Dell, H.R. Morris, D.H. Williams, R.P. Ambler, Biomed. Mass Spectrom. 1 (1974) 269.
[6] S. Hakomori, J. Biochem. (Tokyo) 55 (1964) 205.
[7] H.R. Morris, FEBS Lett. 22 (1972) 257.
[8] H.R. Morris, R.J. Dickinson, D.H. Williams, Biochem. Biophys. Res. Commun. 51 (1973) 247.
[9] H.R. Morris, D.H. Williams, G.C. Midwinter, B.S. Hartley, Biochem. J. 141 (1974) 701.
[10] W.V. Shaw, L.C. Packman, B.D. Burghleigh, A. Dell, H.R. Morris, B.S. Hartley, Nature 282 (1979) 870s.
[11] A. Dell, H.R. Morris, Biomed. Mass Spectrom. 8 (1981) 128.
[12] H.R. Morris, K.E. Batley, N.G.L. Harding, R.A. Bjur, J.G. Dann, R.W. King, Biochem. J. 137 (1974) 409.
[13] K.E. Batley, H.R. Morris, Biochem. Biophys. Res. Commun. 75 (1977) 1010.
[14] K.E. Batley, H.R. Morris, Biochem. Soc. Trans. 5 (1977) 1097.
[15] J. Hughes, T.W. Smith, H.W. Kosterlitz, L. Fothergill, B.A. Morgan, H.R. Morris, Nature 258 (1975) 577.
[16] S. Magnusson, L. Sottrup-Jensen, T.E. Petersen, H.R. Morris, A. Dell, FEBS Lett. 44 (1974) 189.
[17] H.R. Morris, A. Dell, T.E. Petersen, L. Sottrup-Jensen, S. Magnusson, Biochem. J. 153 (1976) 663.
[18] H.R. Morris, M.R. Thompson, A. Dell, Biochem. Biophys. Res. Commun. 62 (1975) 856.

[19] H.C. Thogersen, T.E. Petersen, L. Sottrup-Jensen, S. Magnusson, H.R. Morris, Biochem. J. 175 (1978) 613.
[20] H.R. Morris, M.R. Thompson, D.T. Osuga, A.I. Ahmed, S.M. Chan, J. Vandenheede, R. Feeney, J. Biol. Chem. 253 (1978) 5155.
[21] H.R. Morris, A. Dell, A.E. Banner, A. Evans, R. McDowell, D. Hazelby, Proceedings of the 5th Annual Conference On Mass Spectrometry and Allied Topics, USA, 1977, p. 73.
[22] H.R. Morris, A. Dell, R.A. McDowell, Biomed. Mass Spectrom. 8 (1981) 463.
[23] H.R. Morris, M.R. Thompson, in: A. Frigerio, N. Castagnoli (Eds.), Advances in Mass Spectrometry in Biochemistry and Medicine, vol. 1, Spectrum, New York, 1976, p. 239.
[24] W.T. Shier, T. Lin, A.L. DeVries, Biochem. Biophys. Acta 263 (1972) 406.
[25] G.W. Taylor, H.R. Morris, T.E. Petersen, S. Magnusson, Adv. Mass Spectrom. 18 (1980) 1090.
[26] M. Barber, R.S. Bordoli, R.D. Sedgwick, A.N. Tyler, J. Chem. Soc., Chem. Commun. 7 (1981) 325.
[27] H.R. Morris, M. Panico, G.W. Taylor, Biochem. Biophys. Res. Commun. 117 (1983) 299.
[28] H.R. Morris, M. Panico, M. Barber, R.S. Bordoli, R.D. Sedgwick, A. Tyler, Biochem. Biophys. Res. Commun. 101 (1981) 623.
[29] M. Yamashita, J.B. Fenn, J. Phys. Chem. 88 (1984) 4451.
[30] M.L. Alexandrov, L.N. Gall, N.V. Krasnov, V.I. Nikolaev, V.A. Pavlenko, V.A. Shkurov, Dokl. Akad. Nauk. SSSR 277 (1984) 379.
[31] R.A. McDowell, H.R. Morris, Int. J. Mass Spectrom. Ion Process. 46 (1983) 443.
[32] A. Dell, H.R. Morris, Science 291 (2001) 2351.
[33] R.J. Robb, R.M. Kutny, M. Panico, H.R. Morris, V. Chowdhry, Proc. Natl. Acad. Sci. U.S.A. 81 (1984) 6486.
[34] H. Sasaki, N. Ochi, A. Dell, M. Fukuda, Biochemistry 27 (1988) 8618.
[35] H. Sasaki, B. Bothner, A. Dell, M. Fukuda, J. Biol. Chem. 262 (1987) 12059.
[36] A.L. Chan, H.R. Morris, M. Panico, A.T. Etienne, M.E. Rogers, P. Gaffney, L. Creighton-Kempsford, A. Dell, Glycobiology 1 (1991) 173.
[37] R.A. Siciliano, H.R. Morris, R.A. McDowell, P. Azadi, M.E. Rogers, H.P.J. Bennett, A. Dell, Glycobiology 3 (1993) 225.
[38] A. Dell, H.R. Morris, R.L. Easton, M. Panico, M. Patankar, S. Oehninger, R. Koistinen, H. Koistinen, M. Seppala, G.F. Clark, J. Biol. Chem. 270 (1995) 24116.
[39] H.R. Morris, A. Dell, R.L. Easton, M. Panico, H. Koistinen, R. Koistinen, S. Oehninger, M.S. Patankar, M. Seppala, G.F. Clark, J. Biol. Chem. 271 (1996) 32159.
[40] H.R. Morris, P. Pucci, Biochem. Biophys. Res. Commun. 126 (1985) 1122.
[41] J.S. Cottrell, S. Evans, Anal. Chem. 59 (1987) 1990.
[42] (a) H.R. Morris, R. McDowell, A. Dell, M. Panico, Proceedings of the 39th Conference on Mass Spectrometry and Allied Topics, Nashville, 1991, p. 1693;
(b) H.R. Morris, A. Dell, M. Panico, R. McDowell, Proceedings of the 43rd Conference on Mass Spectrometry and Allied Topics, Atlanta, 1994, p. 360.
[43] J.V. Stone, W. Mordue, K.E. Batley, H.R. Morris, Nature 263 (1976) 207.
[44] H.R. Morris, M. Panico, A. Karplus, P.E. Lloyd, B. Riniker, Nature 300 (1982) 643.
[45] A. Beloff-Chain, J. Morton, S. Dunmore, G.W. Taylor, H.R. Morris, Nature 301 (1983) 255.
[46] H.R. Morris, M. Panico, T. Etienne, J. Tippins, S.I. Girgis, I. MacIntyre, Nature 308 (1984) 746.
[47] D.F. Hunt, J.R. Yates, J. Shabanowitz, S. Winston, C.R. Hauer, Proc. Natl. Acad. Sci. U.S.A. 83 (1986) 6233.
[48] K. Biemann, Meth. Enzymol. 193 (1990) 455.
[49] M.F. Bean, S.A. Carr, G.C. Thorne, M.H. Reilly, S.J. Gaskell, Anal. Chem. 63 (1991) 1473.
[50] K.-H. Khoo, S. Sarda, X.F. Xu, J.P. Caulfield, M.R. McNeil, S.W. Homans, H.R. Morris, A. Dell, J. Biol. Chem. 270 (1995) 17114.
[51] S.M. Haslam, G.C. Coles, E.A. Munn, T.S. Smith, H.F. Smith, H.R. Morris, A. Dell, J. Biol. Chem. 271 (1996) 30561.
[52] D.F. Hunt, R.A. Henderson, J. Shabanowitz, K. Sakaguchi, H. Michel, N. Sevilir, A.L. Cox, E. Appella, V.H. Engelhard, Science 255 (1992) 1261.
[53] D.F. Hunt, H. Michel, T.A. Dickinson, J. Shabanowitz, A.L. Cox, K. Sakaguchi, E. Appella, H.M. Grey, A. Sette, Science 256 (1992) 1817.
[54] J.H.J. Dawson, M. Guilhaus, Rapid Commun. Mass Spectrom. 3 (1989) 155.
[55] H.R. Morris, T. Paxton, A. Dell, J. Langhorne, M. Berg, R.S. Bordoli, J. Hoyes, R.H. Bateman, Rapid Commun. Mass Spectrom. 10 (1996) 889.
[56] P. Teng-umnuay, H.R. Morris, A. Dell, M. Panico, T. Paxton, C.M. West, J. Biol. Chem. 273 (1998) 18242.
[57] S.L. Michell, A.O. Whelan, P.R. Wheeler, M. Panico, R.L. Easton, A. Etienne, S.M. Haslam, A. Dell, H.R. Morris, A.J. Reason, L.J. Herrmann, D.B. Young, G.R. Hewinson, J. Biol. Chem. 278 (2003) 16423.
[58] A.M. Moody, S.J. North, B. Reinhold, S.J. Van Dyken, M.E. Rogers, M. Panico, A. Dell, H.R. Morris, J.D. Marth, E.L. Reinherz, J. Biol. Chem. 278 (2003) 7240.
[59] M. Wacker, D. Linton, P.G. Hitchen, M. Nita-Lazar, S.M. Haslam, S.J. North, M. Panico, H.R. Morris, A. Dell, B. Wren, M. Aebi, Science 298 (2002) 1790.
[60] D. Linton, N. Dorrell, P.G. Hitchen, S. Ames, A.V. Karylshev, H.R. Morris, A. Dell, M.A. Valvano, M. Aebi, B.W. Wren, Mol. Microbiol. 55 (2005) 1695.
[61] M.F. Feldman, M. Wacker, M. Hernandez, P.G. Hitche, C.L. Marolda, M. Kowarik, H.R. Morris, A. Dell, M.A. Valvano, M. Aebi, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 3016.
[62] S. Noguchi, M. Nakano, Biochim. Biophys. Acta 1158 (1993) 217.
[63] G.F. Clark, A. Dell, J. Biol. Chem. 281 (2006) 13853.
[64] R.L. Easton, M.S. Patankar, F.A. Lattanzio, T.H. Leaven, H.R. Morris, G.F. Clark, A. Dell, J. Biol. Chem. 275 (2000) 7731.
[65] E.S. Boja, T. Hoodbhoy, H.M. Fales, J. Dean, J. Biol. Chem. 278 (2003) 34189.
[66] S. Chalabi, M. Panico, M. Sutton-Smith, S.M. Haslam, M.S. Patankar, F.A. Lattanzio, H.R. Morris, G.F. Clark, A. Dell, Biochemistry 45 (2006) 637.
[67] A. Dell, S. Chalabi, R.L. Easton, S.M. Haslam, M. Sutton-Smith, M.S. Patankar, F. Lattanzio, M. Panico, H.R. Morris, G.F. Clark, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 15631.
[68] D. Goldberg, M. Sutton-Smith, J. Paulson, A. Dell, Proteomics 5 (2005) 865.
[69] H. William, H. Press, B.P. Flannery, S.A. Teukolsky, T. William, Vetterling, Numerical Recipes in C: The Art of Scientific Computing, 2nd ed., Cambridge University Press, 1992.
[70] M. Bern, D. Goldberg, J. Comput. Biol. 13 (2006) 364.