Analyst

PAPER

View Article Online
View Journal | View Issue

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

Cite this: *Analyst*, 2015, **140**, 1717

Cluster-based comparison of the peptide mass fingerprint obtained by MALDI-TOF mass spectrometry. A case study: long-term stability of rituximab†

Pablo J. Villacorta,[a] Antonio Salmerón-García,[b] David A. Pelta,[a] José Cabeza,[b] Antonio Lario[c] and Natalia Navas*[d]

We evaluated the use of the peptide mass fingerprint (PMF) obtained by matrix assisted laser desorption and ionization (MALDI) time-of-flight mass spectrometry (TOF-MS) to track changes in the structure of a protein. The first problem we had to overcome was the inherent complexity of the PMF, which makes it difficult to compare. We dealt with this problem by developing a cluster-based comparison algorithm which takes into account the proportional error made by the mass spectrometer. This procedure involves grouping together similar masses in an intelligent manner, so that we can determine which data correspond to the same peptide (any slight differences can be explained as experimental errors), and which of them are too different and thus more likely to represent different peptides. The proposed algorithm was applied to track changes in a commercially available monoclonal antibody (mAb), namely rituximab (RTX), prepared under the usual hospital conditions and stored refrigerated (4 °C) and frozen (−20 °C) for a long term study. PMFs were obtained periodically over three months. For each checked time, five replicates of the PMFs were obtained in order to evaluate the similarities between them by means of the occurrences of the particular peptides (*m*/*z*). After applying the algorithm to the PMF, different approaches were used to analyse the results. Surprisingly, all of them suggested that there were no differences between the two storage conditions tested, *i.e.* the RTX samples were almost equally well preserved when stored refrigerated at 4 °C or frozen at −20 °C. The cluster-based methodology is new in protein mass spectrometry and could be useful as an easy test for major changes in proteins and biopharmaceutics for diverse applications in industry and other fields, and could provide additional stability data in relation to the practical use of anticancer drugs.

Received 7th October 2014,
Accepted 23rd December 2014

DOI: 10.1039/c4an01806k

www.rsc.org/analyst

Introduction

Peptide mass fingerprint (PMF) is based on the enzymatic digestion of proteins and subsequent analysis of the digested peptides by mass spectrometry (MS).[1] PMF is mainly used for protein identification given that each protein undergoes its own specific enzymatic digestion process, which yields a unique set of peptides. The masses of these peptides provide a molecular signature – the fingerprint – that identifies the particular protein.[2] Of the various mass spectrometric techniques that can be used for PMF, matrix-assisted laser desorption and ionization (MALDI) time of flight (TOF) mass spectrometry (MS) is the most commonly used due to its capacity for generating singly charged ions and its relative robustness in the presence of salts and buffers.[3] Although these inherent characteristics make the interpretation of these complex mass spectra easier, sophisticated mathematical and statistical algorithms are essential for analysing and extracting information from the PMF.[4,5]

Mass spectrometric data analysis is a complex procedure which requires a range of signal treatment approaches such as signal smoothing and filtering, detection and/or selection of important features *etc*. A lot of work has been done in this

[a]*CITIC-UGR, Department of Computer Science and Artificial Intelligence, University of Granada, ETSIIT, C/Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain*
[b]*UGC Intercentro Interniveles Farmacia Granada, "San Cecilio Hospital", Biomedical Research Institute ibs.GRANADA. Hospitales Universitarios de Granada, E-18012 Granada, Spain. E-mail: natalia@ugr.es; Fax: +34 958243328; Tel: +34 958243388*
[c]*CSIC (Consejo Superior Investigaciones Científicas), Proteomic Unit, Parasitology and Biomedicine Institute "López Neira", Parque Tecnológico Ciencias de la Salud, Avda. del Conocimiento, s/n. 18100 Armilla, Granada, 18005 Spain*
[d]*Department of Analytical Chemistry, Science Faculty, Biomedical Research Institute ibis.GRANADA, University of Granada, Campus Fuentenueva s/n, E-18071 Granada, Spain. E-mail: natalia@ugr.es; Fax: +34 958243328; Tel: +34 958243388*
†Electronic supplementary information (ESI) available. See DOI: 10.1039/c4an01806k

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

direction,[6,7] including the development of an array of software packages for mass spectrometry analysis that are too numerous to cite here.[3,5,8–12] The authors of ref. 12 have collected and summarized the most important features of many recent open software tools for mass spectrometry analysis on their website; see ref. 13 for an overview.

Although PMFs are mainly used for protein identification, the kind of mass spectrometric data analysis we perform seeks to track the compounds (peptides, in our case) found in successive mass spectra, particularly in PMF, over a period of time. Temporal studies are of great interest for evaluating the evolution of chemical compounds and the implications of any changes detected. The first step in peptide tracking is peptide recognition across different spectra, for which we must return to the general framework of peak matching: in order to follow the evolution of a peptide over a period of time, we must be able to identify the same peptide across several spectra obtained at different moments. We therefore need to find a match between the peaks in the different spectra, so as to be sure that these peaks all correspond to the same peptide. A very common application of this problem is the recognition or classification of substances, for instance bacteria species,[11] which can be identified by the characteristic spectrum they produce.

A great deal of research has been done on the peak alignment problem. Some studies propose a special sample preparation method to ease subsequent manual identification.[3] However, most papers focus on the analysis of the spectra, often using peak alignment algorithms. Many different algorithms have been proposed, especially when TOF-MS is coupled with two dimensional gas (GCxGC-TOF-MS)[8–10,14] or liquid (LCxLC-TOF-MS)[15] chromatography.

The problem we face in our research is slightly different from those cited in the previous paragraph as our data come from matrix-assisted laser desorption/ionization (MALDI) TOF-MS with no chromatographic technique coupled to it. Our aim is to study the temporal evolution of the peptides obtained after the enzymatic digestion of a monoclonal antibody (mAb), namely rituximab (RTX), when prepared and stored under hospital conditions of use, by analysing a sequence of mass spectra (PMFs) obtained at different moments in time from the same vial, *i.e.* immediately after the vial was first opened, after 1 day, 2 days, and so on. We would like to emphasize that several replicated spectra were obtained each day (in exactly the same conditions) in an attempt to assess the reproducibility of the experimental instrumental measurements.

RTX is a chimeric mouse/human IgG1 mAb that binds to CD20, a transmembrane protein, located on pre-B and mature B-lymphocytes.[16] It is intended for use in the treatment of non-Hodgkin's lymphoma,[17] rheumatoid polyarthritis[18] and chronic lymphoid leukemia. mAbs, including RTX, are large glycoproteins (≈150 kDa), composed of four peptide chains, two identical heavy chains (≈50 kDa) and two identical light chains (≈25 kDa) connected by disulfide bonds at their hinge region,[19] which gives them their characteristic Y shape (Fig. 1). Rituximab is a very expensive drug that must be administered sparingly. Our research seeks to contribute to more efficient use of RTX in hospitals by analysing its stability over time. Also, there is a compelling need for additional stability data covering the practical uses of anticancer drugs and adapted

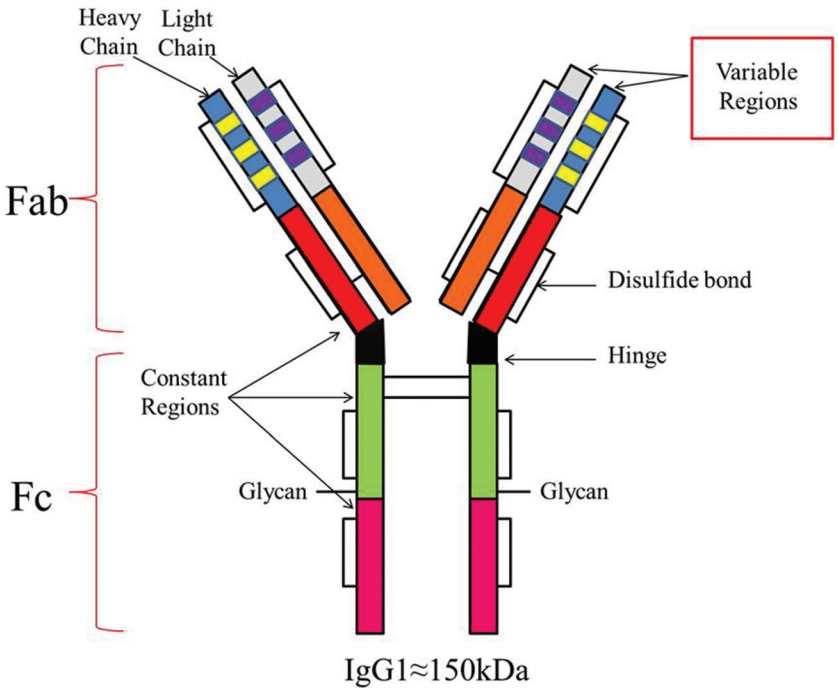

Fig. 1 General structure of IgG1.

This journal is © The Royal Society of Chemistry 2015

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

guidelines for stability studies.[20] In this context, several typical protein characterization methods were recently used in a long-term stability study focusing on the determination of the physicochemical modification of RTX, including size exclusion chromatography (SEC), cation exchange chromatography (CEX), dynamic light scattering (DLS), turbidimetry, second-derivative ultraviolet absorption, Fourier transform infrared spectroscopy (FT-IR) and peptide mapping by HPLC.[21] The results of this interesting work demonstrated that diluted RTX (1 mg mL^{-1} in saline solution) remained stable for at least six months when stored in polyolefin bags at 4 °C. Nevertheless, no research has so far been done using mass spectrometric methodologies. Our research therefore aims to contribute to the in-use stability of this complex drug when prepared in hospital by conducting an in-depth analysis of the mass fingerprint by MALDI-TOF-MS.

As mentioned earlier, the first step for studying the temporal evolution of PMF is the identification of each peptide by its m/z across multiple spectra, which is a peak alignment problem. In general, the alignment methods mentioned above use the bi-dimensional features of chromatography to align the peaks. These methods cannot be used here as we are dealing with MALDI-TOF mass spectra without prior chromatographic separation. Although the intensity values present in the spectra could provide additional information, peptides are mainly characterized in terms of their mass. We have therefore tried to group similar masses together, taking into account the maximum measurement error, which varies depending on the type of mass spectrometer used. In this paper we propose an *ad hoc* fully automated algorithm that exploits the specific tolerance information of the mass spectrometric equipment in an intelligent manner to progressively adjust the mass interval of a peptide and estimate the true peptide masses from samples. This algorithm outputs a more informative summary of the peptides found in the data and was successfully applied to track peptides in the PMF of RTX samples stored under different conditions and checked periodically over three months of a long-term stability study.

MS fingerprint cluster algorithm

Theoretical basis

The fact that we have a set of replicas at each time step poses a challenge in the analysis of the results: can we determine whether two masses from different replicas or days correspond to the same peptide? A correct answer to this question is the first step towards tracking peptides over time. Experimental data of the mass of a given peptide are subject to random errors when using mass fingerprints. The magnitude of the error depends on the resolution of the equipment, but it is assumed that two very close mass measurements correspond to the same peptide. In this study, the idea is to group together peptide masses that are sufficiently similar to allow us to obtain a set of mass measurements for each peptide. We then determine the minimum and the maximum for the set in order to delimit the experimental mass range for the peptide. We also calculate the average for the set, which is an estimation of the true mass of the peptide. From this point onwards, we can define a peptide as a closed interval of real numbers (mass measurements) within which the true mass may lie. We also refer to a peptide with the term *cluster*.

We use a simple additive statistical model for mass observations, similar to that suggested by previous authors when modelling the observed intensity.[22] We assume that each observed mass x' is the sum of the true (unknown) peptide mass x plus (or minus) a mass measurement error $\varepsilon_{\mathrm{ppm}}(x)$. Errors are modelled as independent random variables following a normal distribution $N(0, \sigma_x^2)$ whose variance depends on the true mass x of the peptide. The fact that distribution is normal and centred around 0 means that experimental errors by defect or excess are equally probable and that small errors are more probable than large ones. Mass-dependent variance is due to the equipment we use in our experiments, described in the next section. The manufacturer indicates that the (relative) mass measurement error is 50 ppm, which means that, for example, the error can be up to ±0.05 Da when the true mass being measured is 1000 Da, but it can be up to ±0.15 Da when the mass is 3000 Da. Both assumptions are very common in daily-use equipment, and lead us to conclude that it is more likely that a mass value corresponds to the peptide in the nearest interval than to any other. Modelling experimental errors as independent random variables distributed as $N(0, \sigma_x^2)$ is the key idea behind our mass clustering algorithm.

Clustering algorithms

In an abstract way, dividing a dataset into groups as described in the preceding section, with no prior knowledge of which data should go into each group, is known as clustering. Mass data play the role of *samples*, which are clustered in peptides (mass intervals). Clustering techniques analyse data and divide them into groups of similar samples according to their characteristics. These techniques are applied above all for descriptive purposes, to gain an overview of the data. The distance between the samples can be used to decide whether or not they should be included in the same cluster.

A lot of techniques have been proposed for this task.[23] They can be roughly classified on the basis of two criteria: (a) whether or not they assume a statistical model underlying the data, and (b) whether or not the algorithm needs to know in advance how many clusters it has to make. Perhaps the most famous algorithm is K means,[24] in which the number of clusters K has to be indicated by the user; a lot of variants have also been proposed.

In the case of peptide mass clustering, we apply one-dimensional clustering (only the mass is to be clustered; intensity values are ignored for the moment) in which the number of clusters (peptides) that must be made is not known in advance. Moreover, we assume that the data for each cluster follows a normal distribution centred on the true peptide mass, due to the distribution of experimental errors explained at the beginning of this section. Algorithms that do not

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

require knowledge of the number of clusters in advance are less common; see ISODATA[24] and DBSCAN.[25] However, they cannot be applied directly to our data because of the following constraints C1, C2:

C1: According to the technical characteristics of the mass spectrometer we used, mass measurement errors are up to 50 ppm, which means that two masses that differ by more than 50 ppm must correspond to different peptides. We call it a *cannot-cluster* constraint over these two samples.

C2: In the experimental procedure we followed, a peptide cannot appear more than once in the same replica from the same day. Formally, if we have two measurements $(x_1, y_1)^j_k$ and $(x_2, y_2)^j_k$ that were taken during replica j on day k, with masses x_1 and x_2 and intensities y_1 and y_2, then we can be sure that they correspond to different peptides, no matter how close x_1 and x_2 are. This is another *cannot-cluster* constraint.

Our proposal for a one-dimensional peptide mass clustering algorithm

In order to take the two aforementioned constraints into account, we developed an *ad hoc* constrained-clustering algorithm, which receives an input parameter ppm that stands for the maximum error (in parts per million) caused by the equipment. In our experiments, ppm = 50. Let $\pm\varepsilon_{\text{ppm}}(x)$ be the maximum error introduced when measuring mass x under such conditions, where $\varepsilon_{\text{ppm}}(x) = \text{ppm} \times 10^{-6} \times x$. Note that whilst every sample $(x_i, y_i)^j_k$ considered during the algorithm contains information about the replica j and the day k when it was taken, only the mass information x_i is used for clustering.

We use the following notation (Fig. 2). Let D be the number of different days in which we have taken measurements, r the number of replicas performed every day, and n_{jk} the number of samples that were obtained in replica j on day k. Letters C_t refer to a cluster, *i.e.* a set of masses already clustered together, and we denote $[C_t^{\min}\ C_t^{\max}]$ the corresponding mass interval defined by the minimum and maximum values of C_t. Let avg_t be the average value of masses within cluster C_t, which we use as an estimate of the true mass. The maximum amplitude of such a cluster would be $[\text{avg}_t - \varepsilon_{\text{ppm}}(\text{avg}_t), \text{avg}_t + \varepsilon_{\text{ppm}}(\text{avg}_t)]$. The algorithm proceeds as follows.

Algorithm: constrained peptide mass clustering

INPUTS: ppm $\in \mathrm{R}^+$ (needed to compute $\varepsilon_{\text{ppm}}(x) = \text{ppm} \times 10^{-6} \times x$)

$D \times r$ sets of mass-intensity samples $\{(x_i, y_i)^j_k\}$; $i = 1,\ldots, n_{jk}$; $j = 1,\ldots, r$; $k = 1,\ldots, D$

OUTPUT: set of clusters C_t with their corresponding mass intervals $[C_t^{\min}, C_t^{\max}]$ and avg_t.

Stage 1: generate the initial set of clusters.

Take the samples $\{(x_i, y_i)^1_0\}$, $i = 1,\ldots, n_{1,0}$, *i.e.*, the data from replica 1 from day 0, and for each x_i create a cluster with that sample. For now, the average value of every cluster matches the (only) point included in it. The output of this stage is a set of $n_{1,0}$ clusters C_t.

Stage 2: map samples to intervals.

For each replica j of each day k ($j \neq 1$ when $k = 0$):

For each point $(x_i, y_i)^j_k$ in that replica:

Find C_t: $C_t^{\min} \leq x_i \leq C_t^{\max}$.

If C_t exists **and** $\nexists\ x_s$: $(x_s, y_s)^j_k \in C_t$ // (checking C2)

Do $C_t \leftarrow C_t \cup \{(x_i, y_i)^j_k\}$ and update avg_t

Call CheckSurrounding (C_t) to propagate the effect of updating avg_t.

If C_t exists **and** $\exists\ x_s$: $(x_s, y_s)^j_k \in C_t$, then x_i and x_s must not be clustered together as they come from distinct peptides: // (checking C2)

Create a new cluster C_x with $(x_i, y_i)^j_k$, and set $\text{avg}_x \leftarrow x_i$.

Split the cluster C_t found by calling CheckSurrounding (C_x) to redistribute its elements between C_t and the new C_x.

The rest of the surrounding clusters that may be affected by the previous step will have been updated by the previous CheckSurrounding call.

If C_t does not exist,

Find the nearest existing cluster C_{t^*} where $t^* = \text{argmin}_t\{|x_i - \text{avg}_t|\}$.

Let $d^* \leftarrow |x_i - \text{avg}_{t^*}|$.

If $d^* \leq \varepsilon_{\text{ppm}}(\text{avg}_{t^*})$ // (checking C2)

Do $C_{t^*} \leftarrow C_{t^*} \cup \{(x_i, y_i)^j_k\}$ and update avg_{t^*}, $C_{t^*}^{\min}$ or $C_{t^*}^{\max}$

Call CheckSurrounding (C_{t^*}) to propagate the updating of avg_{t^*}

If $d^* > \varepsilon_{\text{ppm}}(\text{avg}_{t^*})$ // (checking C2)

Create a new cluster C_x with $(x_i, y_i)^j_k$, and set $\text{avg}_x \leftarrow x_i$

Call CheckSurrounding (C_x)

Stage 3.

For each cluster, try to merge it with the surrounding ones by checking that no pair of points from the same replica will be clustered together, and that the distance from the extremes of the merged interval to the updated average mass $\text{avg}_{\text{merged}}$ is less than or equal to $\varepsilon_{\text{ppm}}(\text{avg}_{\text{merged}})$.

Procedure CheckSurrounding (INPUT: a cluster C_t)

For all C_x such that $|\text{avg}_x - \text{avg}_t| \leq 1$ Da // propagate to surrounding peptides only (faster)

For all $(x, y)^j_k \in C_x$

If $|x - \text{avg}_t| \leq |x - \text{avg}_x|$ **and** $\nexists\ x_s$: $(x_s, y_s)^j_k \in C_t$, move x to C_t: // (checking C1)

Do $C_t \leftarrow C_t \cup \{(x, y)^j_k\}$ and update avg_t, $C_t^{\min}$ or $C_t^{\max}$

Do $C_x \leftarrow C_x \backslash \{(x, y)^j_k\}$ and update avg_x, $C_x^{\min}$ or $C_x^{\max}$

End

 This journal is © The Royal Society of Chemistry 2015

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

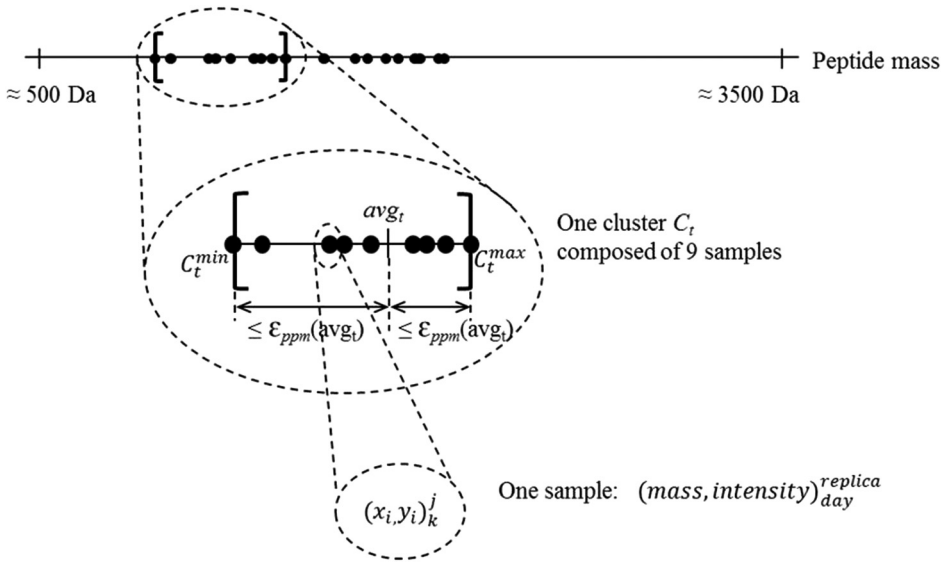

Fig. 2 Mass clustering scheme proposed.

The condition $\nexists\, x_s$: $(x_s, y_s)_k^j$ amounts to saying that no other mass coming from the same replica from the same day had been previously added to the same cluster.

The clusters obtained after Stage 3 allow us to count the number of replicas from the same day in which a given peptide (represented by a cluster, say C_t) occurs, *i.e.* whether the peptide has been detected in 1, 2,... up to 5 replicas on a given day of interest, say $k = k_0$. This can be done by counting the number of samples with the form $(x_i, y_i)_{k_0}^j$ that belong to the cluster C_t, which represents the peptide being studied, as explained in the next section. It should be noted that it would be almost impossible to make this count accurately without the aid of our clustering algorithm.

Experimental

Substances and solvents

All reagents were of analytical reagent grade unless otherwise stated. Reverse-osmosis quality water (purified with a Milli-RO plus Milli-Q station from Millipore Corp., Madrid, Spain) was used throughout. Trifluoroacetic acid (TFA) was from Merk KGaA (Darmstadt, Germany) and acetonitrile from Poch S.A (Gliwice, Poland). The isotonic solution of 0.9% NaCl was supplied by B. Braun Medical (Madrid, Spain). Ammonium bicarbonate, dithiotheitol (DTT), iodoacetamide and α-cyano-4-hydroxycinnamic acid (α-CHCA) were supplied by Sigma-Aldrich (Barcelona, Spain). Trypsine Gold (Mass Spectrometry Grade) was from Promega Corporation (Madrid, Spain).

Rituximab solutions

RTX solutions of 1.0 mg mL^{-1} were prepared from the authorized medicine Mabthera® (Roche Pharma AG, Grenzach-Wyhlen, Germany). The medicine indicates a quantitative composition of 100 mg of RTX in each single-use glass vial with 7.35 mg mL^{-1} sodium citrate dihydrate, 9 mg mL^{-1} sodium chloride, sodium hydroxide and hydrochloric acid to obtain a pH of 6.5, and polysorbate 80 (PS80) as a stabilizing agent. This concentrated solution of 10 mg mL^{-1} of RTX is ready to dilute in 0.9% sodium chloride solution, according to the manufacturer's instructions.[26] A single vial was used to prepared RTX solutions of 1.0 mg mL^{-1}.

Long-term study

A sample of 1.0 mg mL^{-1} in 0.9% NaCl was prepared from the medicine Mabthera®. One aliquot was stored refrigerated at 4 °C and protected from daylight, and several aliquots were stored frozen at −20 °C. In the long-term study, samples from the two storage conditions were analysed on day 0 (control day), 1, 3, 4, 7, 14, 28, 44, 58, 73 and 88 (three months).

Enzymatic digestion and MALDI-TOF-MS analysis

Enzymatic digestion. RTX solution of 1.0 mg mL^{-1} was diluted with ammonium bicarbonate 50 mM to end up with 3 μg of protein. The reduction/alkylation of the disulfide bonds prior to trypsin addition was performed by adding DTT 10 mM solution to the diluted RTX with ammonium bicarbonate 50 mM and this mixture solution was incubated at 55 °C for 60 min (reduction step); after that, iodoacetamide solution (43 mM prepared in ammonium bicarbonate 50 mM) was added and the mixture was incubated at room temperature for 30 min in the dark (alkylation step). The trypsin digestion process was as follows: 150 ng of trypsin gold was added to the previous solution, which was then incubated at 37 °C for 4 hours. To stop the digestion process, we added sufficient TFA to reach a volume of 0.2%.

MALDI-TOF-MS analysis. 1 μL of the RTX digested sample was mixed with 1 μL of the α-cyano matrix solution (α-CHCA 5 mg ml^{-1}, 50% acetonitrile and 0.1% TFA) and 1 μL was loaded on the stainless steel MALDI target and dried in air.

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

Table 1 Masses (*m*/*z*, in Da) to be removed from data files (PMF)

515.33	559.29	679.51	864.49	1234.67	1493.75	2225.12
524.16	568.13	701.49	870.54	1265.63	1707.78	2233.00
534.18	570.70	823.11	892.50	1300.58	1716.85	2383.95
537.31	590.11	825.10	1037.55	1365.64	1940.93	2717.05
546.15	634.07	842.50	1045.56	1383.69	1765.73	
548.19	650.05	845.10	1126.56	1434.77	2082.98	
550.16	656.05	856.52	1179.60	1475.78	2211.10	

The mass spectra were acquired by a Voyager DE-PRO (Applied Biosystems) MALDI-TOF mass spectrometer equipped with a standard nitrogen laser (337 nm) in positive reflectron mode. At least 200 laser shots were collected for each spectrum and analyzed using Voyager™ 5 Software (Applied Biosystems). A mass spectrum was taken for each drop between 0 and 5000 *m*/*z*. The recalibration of the mass spectrometer with a commercial standard of peptide mixture (Pepmix, Bruker) was performed every 20 mass spectra recorded.

The mMass 5.5.0 program (available free online) was also used throughout the study to manage the mass spectrometric data. All the mass spectrometric figures were made with this program.[12,27,28]

MS fingerprint data pre-processing for clustering analysis

Before exporting to a text file for the subsequent clustering analysis, each spectrum was processed as follows:

We applied a Peak Detection with threshold = 0, so no value is excluded.

We applied an Advanced Baseline Correction with Peak width = 32, Flexibility = 0.5 and Degree = 0.1.

A smoothing was carried out using correlation factor = 0.7.

We applied a linear calibration using 3 known masses of RTX: 838.50 Da, 1808.00 Da and 3334.64 Da.

We applied a deisotoping algorithm, Voyager™ 5 Software (Applied Biosystems).

Since every file still contains about 3300 pairs (mass, intensity), most of which are just noise because the peak detection threshold was set to 0, we decided to retain only the top 200 data pairs with the highest intensity in each file.

As a result, we obtained 60 text files, each containing 200 pairs of data (mass, intensity).

When any of the masses set out in Table 1 was found in the files, it was removed because these masses do not correspond to RTX peptides but were originated by other external substances used in the experiment.

Results and discussion

RTX mass fingerprint

Another important aspect of this research is the replication of the measurements within the same day. Although the five replicates were acquired successively from different spots on the same sample, a number of inherent conditions of the technique (MALDI-TOF-MS) caused significant variability in the PMF. The entire signal was considered and used for the peak detection procedure before applying the proposed algorithm as indicated in the Experimental section. As expected, the number of *m*/*z* data was different for each one of the five replicates. For example, for the control sample (day 0), the *m*/*z* data were: 3346 (replicate 1), 3326 (replicate 2), 3430 (replicate 3), 3415 (replicate 4) and 3250 (replicate 5). We decided not to look for minimal changes in the mAb structure due to the high level of noise in the PMF, *i.e.* the large number of signals obtained (almost 3500 data in each mass spectra) with low intensity values (mainly noise) and the difficulty of assessing significant differences between mass spectra. The limited resolution of the equipment was also a constraint. We therefore decided to focus on major changes considering also that in a long-term study major structural modifications are also likely. We are not identifying the particular residues where the modification/alterations are produced. Thus the mass spectra were simplified to the 200 most intense peaks. This simplification makes sense if we take into account the RTX primary sequence published in ref. 29–31 (ESI, Fig. 1†). The theoretical digestion by trypsin yields 89 peptides when we allow up to one missed cleavage (ESI, Table 1†). Therefore 200 peaks comfortably cover this value of 89 theoretical peptides from RTX.

The hypothesis to support the differential peptide pattern as fingerprinting for stability and its correlation with the antibody stability is: trypsin is a very specific protease that cleaves at arginine and/or lysine residues, therefore if changes/modifications occur in these residues the pattern of the trypsin cleavage will be different because of the missing cleavages. It is interesting to note that deamidation is one of the most important modifications that occur in proteins, and both amino acid residues involved in the trypsin cleavage (arginine and lysine) suffer from this type of modification in their lateral chain. Therefore, only considering this modification, it could be inferred that the PMF after deamidation will be different from the initial protein structure. In brief, if modifications occur in the protein structure, the pattern of cleavage would be different (concretely if the modifications affect arginine and lysine as they are the cleavage points), therefore the PMF will also be different.

Nevertheless, even using MS reduced to 200 peaks, comparison between replicates, or over time or between different storage conditions was difficult. The RTX PMF of replicates 1 and 5 from samples at checked day 0 (control day), day 44 (middle checked time) and day 88 (last checked day) stored refrigerated at 4 °C are shown in Fig. 3. In this figure we can see that mass spectra replicates are not exactly the same and that the 200 most intense peaks vary from one spectrum to the next. Visual comparison of the mass spectra from the entire long-term study (60 PMFs for each stored condition studied) to evaluate changes in the RTX structure is unfeasible in practice, since the inherent characteristics of the techniques produce similar, but not exactly the same, values for the *m*/*z* of each particular peptide, and manual tracking of every single peptide is impractical. We therefore decided to use the pro-

 This journal is © The Royal Society of Chemistry 2015

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

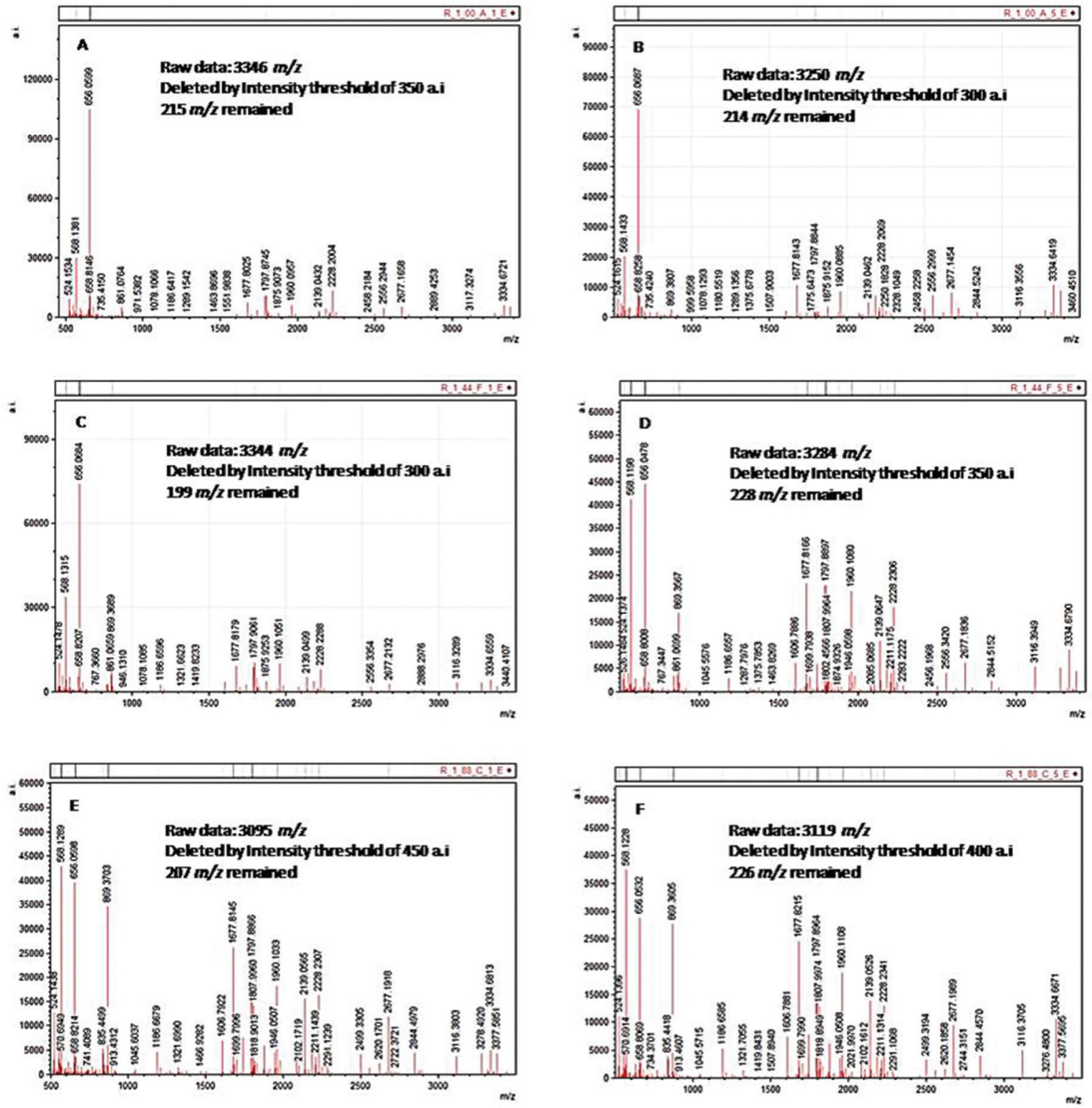

Fig. 3 RTX PMF after *m/z* filtering. A and B: replicates 1 and 5 from day 0 (control day). C and D: replicates 1 and 5 from day 44 (middle checked time). E and F: replicates 1 and 5 from day 88 (last checked day).

posed algorithm for tracking particular peptides over the study period (three months). The results are discussed in the next section.

In each mass spectrum the intensity of the peaks is affected by a specific offset which all the intensity data of a file increase or decrease. Although this aspect is expected, it must be taken into account when assessing the temporal evolution of peptide abundance, using the relative data for intensity rather than the absolute.

Application of the peptide mass clustering algorithm to RTX mass fingerprint

The outcomes of the application of the proposed peptide cluster algorithm to the entire data set (60 files) allow us to track them over time for each storage condition (refrigerated and frozen). Once the peptides are assigned to the specific clusters by the algorithm, the most reproducible ones are defined as those identified in both the five PFM replicates from the same checked day and the 100 most intense signals. Although the average peptide mass was calculated using the 200 most intense data, when it came to analysing the most characteristic peptides from the protein for each checked day, we used only the top 100. These most reproducible peptides can be considered the most representative of the particular protein. Therefore, in the case of RTX, the most representative peptides for defining the fresh protein (unchanged) are those identified in all five replicates and amongst the 100 most intense in the PMF for the control day (day 0), on which no

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

RTX modifications are expected since the samples were prepared and analysed immediately after opening the medicine. The evolution of these representative peptides over the allotted time period (three months) can be tracked in order to evaluate whether RTX undergoes any changes during storage. Any changes in the RTX structure are registered by these representative peptides by a change in the number of occurrences among the 100 most intense signals in the following checked days.

Table 2 RTX representative peptides tracked over time by occurrences. Refrigerated sample

Mass lower bound	Mass upper bound	Average *m/z* 1 (peptide)	Day 0 (control)	Day 1	Day 3	Day 4	Day 7	Day 14	Day 21	Day 28	Day 44	Day 58	Day 73	Day 88	Occurrences[2]
522.117	522.153	522.135	5	5	5	5	5	5	5	5	5	5	5	5	60
526.146	526.182	526.161	5	5	5	5	5	5	5	5	5	5	5	5	60
566.092	566.144	566.116	5	5	5	5	5	5	5	5	5	5	5	5	60
570.128	570.166	570.143	5	5	5	5	3	5	5	5	5	5	5	5	58
594.898	594.966	594.924	5	5	5	5	5	4	5	4	5	5	5	5	58
630.039	630.082	630.057	5	5	5	5	5	5	5	5	5	5	3	3	56
637.276	637.354	637.308	5	0	0	0	0	0	0	0	0	0	0	0	5
647.318	647.37	647.352	5	1	4	0	0	0	0	1	1	1	0	0	13
658.786	658.845	658.816	5	5	5	5	5	5	5	5	5	5	5	5	60
659.223	659.303	659.276	5	0	0	0	1	0	0	0	1	0	0	0	7
678.023	678.079	678.054	5	5	5	5	5	5	5	5	5	5	5	5	60
691.342	691.401	691.385	5	0	4	0	0	0	0	0	0	0	0	0	9
700.005	700.069	700.04	5	5	5	4	5	3	4	4	5	5	5	4	54
703.414	703.507	703.468	5	0	5	0	4	0	4	0	2	0	0	0	20
735.398	735.467	735.418	5	0	4	0	1	0	0	0	0	0	0	0	10
779.432	779.466	779.445	5	0	4	0	0	0	0	0	0	0	0	0	9
801.059	801.149	801.119	5	4	5	4	1	2	4	1	3	3	4	3	39
823.436	823.532	823.476	5	0	3	0	0	0	0	0	0	0	0	0	8
838.495	838.521	**838.504**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
861.06	861.11	861.081	5	5	5	5	5	3	5	4	5	5	5	5	57
867.061	867.116	867.085	5	5	5	5	5	4	5	4	5	5	5	5	58
869.351	869.385	869.369	5	5	5	5	5	5	5	5	5	5	5	5	60
1078.09	1078.15	1078.12	5	3	5	4	5	1	4	0	1	4	5	1	38
1606.77	1606.85	**1606.8**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
1677.8	1677.84	**1677.82**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
1699.78	1699.85	1699.81	5	5	5	5	5	5	5	5	5	5	4	5	59
1740.8	1740.9	**1740.87**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
1791.8	1791.83	**1791.82**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
1797.87	1797.91	1797.89	5	5	5	5	5	5	5	5	5	5	5	5	60
1807.96	1808.02	**1808.00**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
1813.75	1813.82	1813.79	5	5	5	5	5	5	5	5	5	5	5	5	60
1818.85	1818.93	**1818.9**	**5**	**5**	**5**	**5**	**4**	**5**	**5**	**2**	**5**	**5**	**5**	**5**	**56**
1875.91	1875.97	1875.93	5	5	5	5	5	5	5	5	5	5	5	5	60
1946.01	1946.06	**1946.05**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**2**	**5**	**57**
1960.09	1960.15	1960.11	5	5	5	5	5	5	5	5	5	5	5	5	60
1982.05	1982.14	1982.09	5	5	5	5	5	5	5	5	5	5	5	5	60
2082	2082.06	**2082.03**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**3**	**5**	**58**
2139.04	2139.08	2139.05	5	5	5	5	5	5	5	5	5	5	5	5	60
2141.07	2141.09	2141.08	5	0	5	0	2	0	5	0	5	5	5	0	32
2183.05	2183.1	**2183.07**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
2204.99	2205.07	2205.04	5	5	5	5	5	5	5	5	5	5	5	5	60
2228.2	2228.25	**2228.23**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
2250.13	2250.24	2250.19	5	5	5	5	5	5	5	5	5	5	5	5	60
2283.17	2283.3	2283.21	5	2	5	3	4	3	5	3	4	4	1	5	44
2499.23	2499.36	2499.3	5	5	5	2	4	3	5	0	5	5	5	5	49
2556.29	2556.43	2556.33	5	5	5	5	5	5	5	5	5	5	5	5	60
2620.12	2620.23	**2620.18**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**1**	**0**	**5**	**51**
2677.14	2677.23	**2677.19**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
2722.27	2722.4	**2722.33**	**5**	**1**	**5**	**0**	**1**	**0**	**5**	**0**	**4**	**5**	**5**	**2**	**33**
2844.43	2844.6	2844.49	5	5	4	5	5	5	3	5	5	1	0	5	48
3278.4	3278.55	3278.47	5	5	5	5	5	5	5	5	5	5	5	5	60
3320.45	3320.6	**3320.52**	**5**	**4**	**5**	**1**	**2**	**4**	**4**	**0**	**4**	**5**	**5**	**5**	**44**
3334.63	3334.72	3334.66	5	5	5	5	5	5	5	5	5	5	5	5	60
3377.48	3377.62	3377.56	5	5	5	5	5	5	5	5	5	5	5	5	60
666.004	666.063	666.035	4	0	3	0	5	0	0	0	0	0	0	0	12
676.406	676.466	676.437	4	2	2	1	0	0	1	1	1	2	0	1	15
911.026	911.126	911.071	4	2	3	2	5	0	3	0	1	4	4	1	29
1829.93	1830.06	1829.98	4	5	5	5	5	5	5	5	5	3	4	4	55
2889.27	2889.43	**2889.35**	**4**	**5**	**5**	**5**	**4**	**4**	**5**	**5**	**1**	**5**	**0**	**1**	**44**
3116.31	3116.5	3116.38	4	5	5	5	5	5	5	5	5	5	5	5	59

This journal is © The Royal Society of Chemistry 2015

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

Although the algorithm was applied separately to the two storage conditions, the results for the control day for the most representative peptides were identical. This result validates the algorithm for its intended purpose considering that the input data were different for each storage condition, sharing only the five PMF replicates for the control day. The number of representative peptides for the control day (day 0) was 60; 54 of which were detected in all five PMF replicates and 6 in four replicates. 15 of these peptides could be explained by RTX theoretical enzymatic digestion (Table 2). Tables 2 and 3 show these representative peptides, and they are also represented in Fig. 4.

Table 3 RTX representative peptides tracked over time by occurrences. Frozen sample

Mass lower bound	Mass upper bound	*m/z* 1 (peptide)	Day 0 (control)	Day 1	Day 3	Day 4	Day 7	Day 14	Day 21	Day 28	Day 44	Day 58	Day 73	Day 88	Occurrences[2]
522.122	522.158	522.136	5	5	5	5	5	5	5	5	5	5	5	5	60
526.147	526.191	526.162	5	5	5	5	5	5	5	5	5	5	5	5	60
566.095	566.15	566.118	5	5	5	5	5	5	5	5	5	5	5	5	60
570.127	570.169	570.142	5	5	4	4	4	5	5	5	5	5	5	4	56
594.905	594.963	594.926	5	5	5	5	5	5	5	5	5	5	5	5	60
630.04	630.084	630.057	5	5	5	5	5	5	4	5	5	5	5	5	59
637.29	637.324	637.302	5	0	0	0	0	0	1	0	0	0	0	0	6
647.327	647.397	647.362	5	0	5	0	1	0	2	1	0	0	0	1	15
658.794	658.848	658.816	5	5	5	5	5	5	5	5	5	5	5	5	60
659.23	659.304	659.276	5	0	0	0	0	0	0	0	4	0	0	0	9
678.029	678.097	678.056	5	5	5	5	5	5	5	5	5	5	5	5	60
691.36	691.414	691.386	5	0	4	0	2	0	1	0	0	0	0	0	12
699.997	700.075	700.04	5	5	5	5	4	4	5	4	4	5	5	2	53
703.422	703.5	703.467	5	0	5	0	5	0	3	0	3	0	0	0	21
735.323	735.43	735.398	5	0	3	0	1	0	1	0	0	0	0	0	10
779.411	779.473	779.44	5	0	1	0	0	0	0	0	0	0	0	0	6
801.087	801.146	801.122	5	4	1	3	5	3	4	2	4	5	5	0	41
838.498	838.521	**838.505**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**4**	**5**	**5**	**5**	**5**	**59**
861.051	861.108	861.078	5	5	5	5	5	5	5	5	5	5	5	5	60
867.062	867.113	867.085	5	5	5	5	5	5	5	5	5	5	5	5	60
869.353	869.395	869.369	5	5	5	5	5	5	5	5	5	5	5	5	60
1078.08	1078.14	1078.11	5	2	2	4	5	2	3	3	1	5	4	0	36
1606.77	1606.85	**1606.8**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
1677.8	1677.84	**1677.82**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
1699.78	1699.84	1699.81	5	5	5	5	5	5	5	4	5	5	1	5	55
1740.83	1740.9	**1740.87**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
1791.8	1791.83	**1791.82**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
1797.87	1797.91	1797.89	5	5	5	5	5	5	5	5	5	5	5	5	60
1807.98	1808	**1808.00**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
1813.72	1813.82	1813.78	5	5	5	5	5	5	5	4	5	5	5	5	59
1818.87	1818.94	**1818.9**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**4**	**5**	**5**	**5**	**5**	**59**
1875.91	1875.94	1875.93	5	5	5	5	5	5	5	5	5	5	5	5	60
1946	1946.08	**1946.04**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**2**	**5**	**57**
1960.09	1960.13	1960.11	5	5	5	5	5	5	5	5	5	5	5	5	60
1982.06	1982.12	1982.09	5	5	5	5	5	5	5	4	5	5	5	5	59
2082.01	2082.07	**2082.03**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**4**	**5**	**59**
2139.03	2139.08	2139.05	5	5	5	5	5	5	5	5	5	5	5	5	60
2141.07	2141.1	2141.08	5	1	5	0	1	0	5	2	4	5	5	0	33
2183.05	2183.11	**2183.07**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
2204.99	2205.07	2205.03	5	5	5	5	5	5	5	4	5	5	4	5	58
2228.2	2228.25	**2228.23**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
2250.16	2250.24	2250.2	5	5	5	5	5	5	5	5	5	5	5	5	60
2283.12	2283.27	2283.2	5	1	5	2	5	4	5	5	4	3	1	5	45
2499.22	2499.36	2499.31	5	5	5	5	5	5	5	5	4	5	5	5	59
2556.29	2556.38	2556.33	5	5	5	5	5	5	5	5	5	5	5	4	59
2620.11	2620.27	**2620.18**	**5**	**5**	**4**	**5**	**5**	**5**	**5**	**5**	**5**	**1**	**1**	**5**	**51**
2677.15	2677.22	**2677.19**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**60**
2722.27	2722.4	**2722.33**	**5**	**4**	**5**	**0**	**3**	**1**	**5**	**3**	**1**	**5**	**5**	**0**	**37**
2844.41	2844.55	2844.49	5	5	5	5	5	5	5	5	5	0	0	5	50
3278.39	3278.56	3278.46	5	5	5	5	5	5	5	5	5	5	5	5	60
3320.45	3320.61	**3320.52**	**5**	**4**	**4**	**2**	**3**	**3**	**5**	**4**	**4**	**5**	**3**	**2**	**44**
3334.64	3334.7	3334.66	5	5	5	5	5	5	5	5	5	5	5	5	60
3377.46	3377.61	3377.55	5	5	5	5	5	5	5	5	5	5	5	5	60
666.004	666.089	666.038	4	0	3	0	4	0	0	0	0	0	0	0	11
911.013	911.132	911.068	4	3	3	1	4	2	2	0	3	4	5	0	31
1829.92	1830.02	1829.97	4	5	4	5	5	5	5	4	5	5	1	5	53
2889.27	2889.43	**2889.34**	**4**	**4**	**4**	**5**	**3**	**5**	**4**	**5**	**2**	**5**	**0**	**2**	**43**
3116.29	3116.46	3116.37	4	5	5	5	5	5	5	5	5	5	5	5	59

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

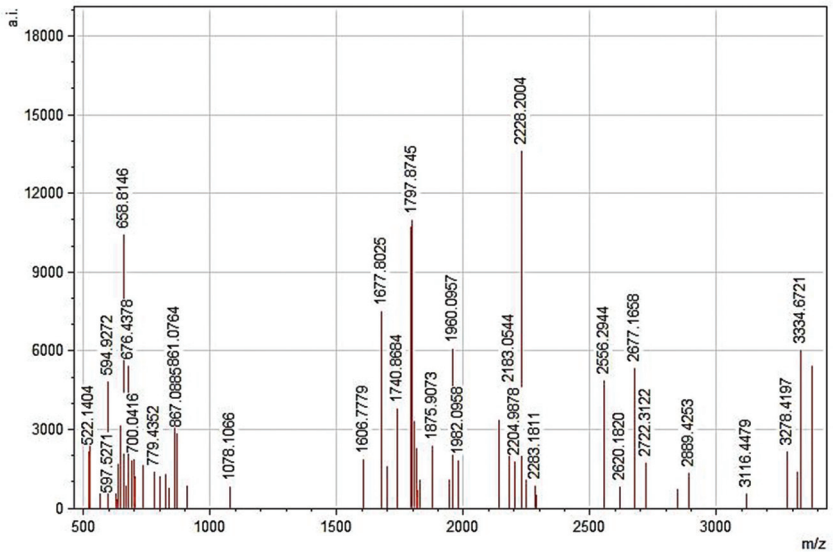

Fig. 4 Most representative peptides of RTX (control day).

The next step was to track these representative peptides over time. Tables 2 and 3 summarize the results for the RTX samples stored refrigerated at 4 °C and frozen at −20 °C respectively. In the first three columns of these tables, the most representative peptides (*m*/*z*) for fresh RTX (those with five and four occurrences amongst the 100 most intense signals in the PMF after applying the algorithm) are indicated. The first and second columns show the lower and upper bounds of the mass interval (cluster) created by the algorithm for a specific peptide, and the third contains the average mass for all the data included in that peptide by the algorithm. It is important to notice how narrow these intervals are despite containing up to 60 samples each. Moreover, since experimental error grows with the magnitude of the true mass being measured, the intervals tend to be slightly wider when the masses they cover are larger.

The evolution of the peptides over time can be tracked in the rows of Tables 2 and 3, where the number of occurrences for each day is specified. The maximum number of occurrences is 60 (5 replicates per 12 checked days). Those peptides with more than 55 occurrences are indicated in shaded grey. Peptides with the maximum number of occurrences are distinguished using dark shaded grey. When the number of occurrences was below 60 but the missed occurrences were randomly distributed over time, peptides are highlighted in light shaded grey. We assumed in this case that the peptide was not detected in the particular PMF due to a random error. Taking this into account, the percentage of the peptides that were conserved over the three months was unexpectedly similar for the two storage conditions, at about 60%. Of the 60 most intense peptides detected on the control day (day 0), 37 were detected at the end of the three month study period (day 88) for the sample stored refrigerated and 38 for the sample stored frozen. These peptides were almost the same for both storage conditions, differing only slightly in peptides 630.05 *m*/*z*, 2499.31 *m*/*z* (most conserved when frozen). Surprisingly, the results for both storage conditions were very similar, not only in the peptides that were conserved, but also in those not conserved and in their evolution over time (number of occurrences). Although there were several results that were difficult to explain, such as 0 occurrences for one day followed by 5 occurrences on the next checked day (*i.e.* 2282.2 *m*/*z*), this happened rarely and could be accepted within the context of the complexity of the problem. In general, the evolution of the peptides was consistent without random fluctuations.

Regarding those peptides not conserved over time, it is also interesting to note that the main changes occurred after the first 24 hours of storage (day 1). Eight of these peptides were not detected from the first day onwards, a result that was the same for both storage conditions, *i.e.* 637.308 *m*/*z*, 659.276 *m*/*z*, 691.385 *m*/*z*, 703.468 *m*/*z*, 735.418 *m*/*z*, 779.445 *m*/*z*, 823.476 *m*/*z*, 666.035 *m*/*z*. The peptides with 2141.08 *m*/*z* and 2283.21 *m*/*z* were also not detected at day 1 but there was no pattern to their evolution over time. For the rest of the peptides, although their evolution did not show a clear pattern, it can be assumed that changes did occur because they were not entirely conserved over the study period.

The results obtained after applying the algorithm can be examined in a different way by analysing the most representative peptides for each checked day. The goal is to track those peptides that are detected as new. In this case we compared the results for the mass spectra of these representative peptides for each day with the mass spectra of the most representative peptides for the control day, day 0. The results are summarized in Fig. 5 and 6. For each checked day the number of representative peptides was different, but always between 60 and 72 peptides. Again, similar results were obtained for both storage conditions (see Fig. 5 and 6) as deduced from the same patterns in the mass spectra. This method of analysis

 This journal is © The Royal Society of Chemistry 2015

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

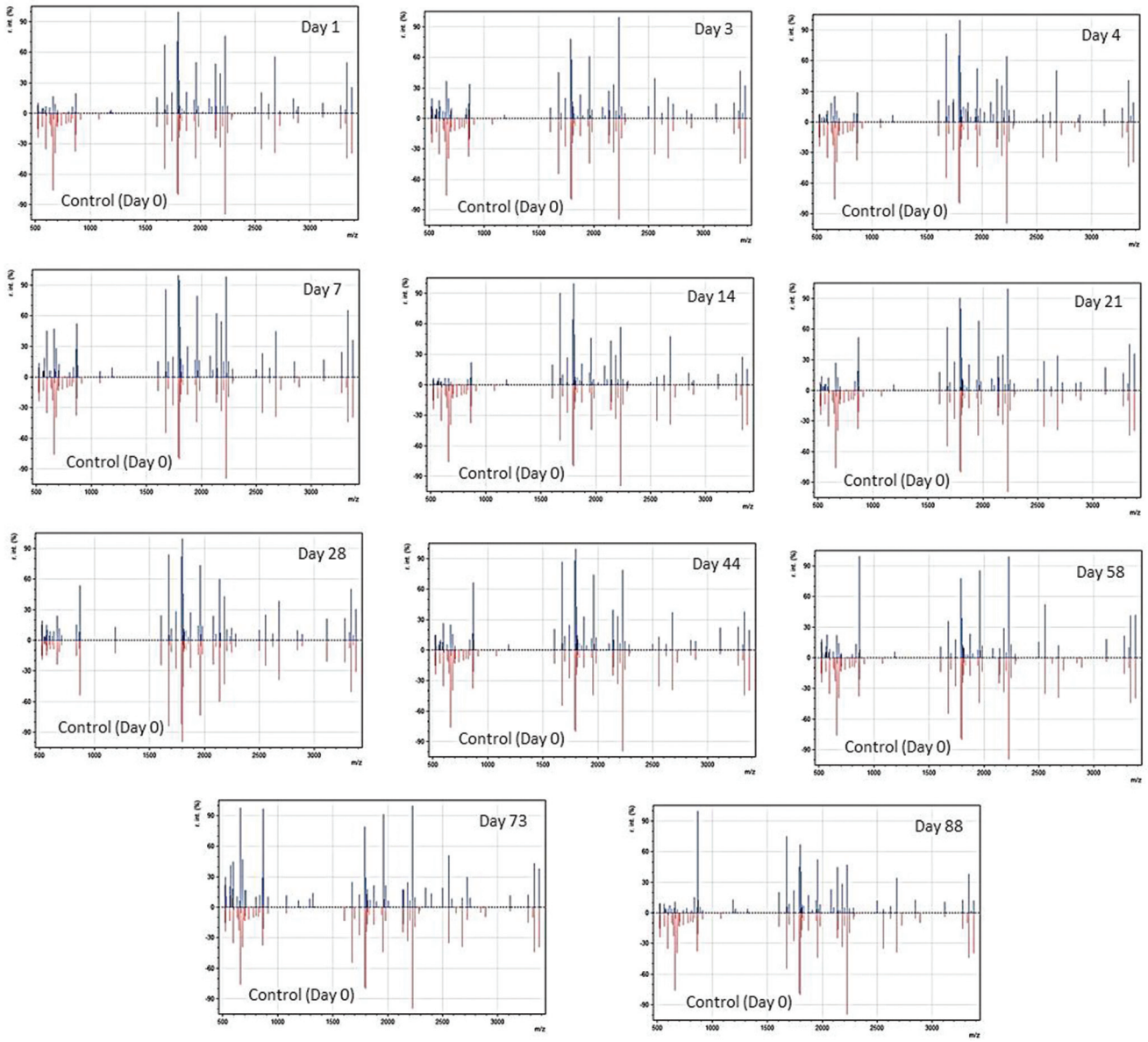

Fig. 5 Mass spectra of the most representative peptide over time *versus* control day. Refrigerated samples.

also allowed for graphical evaluation of signal intensities. In particular, we noted changes in the intensities over time, which could be related to changes in the different forms (isoforms or heterogeneity) of the native RTX. On the other hand, an overall analysis of the main peaks for the control day over time suggested the absence of important damage in the primary structure of the RTX since they were detected every checked day and in high percentages. For example, the main peaks for the control day at 2228.20 *m*/*z* (100%, base peak), 1797.87 *m*/*z* (81%), 1791.80 *m*/*z* (79%), 1677.80 *m*/*z* (55%) were always detected under both storage conditions. An intensity (in %) *versus* time plot did not indicate a clear trend, but the main peaks were consistently around 50% of the relative intensity. It is important to note that the peaks at 2228.20 *m*/*z*, 1791.80 *m*/*z* and 1677.80 *m*/*z* can be explained by the theoretical digestion of the RTX, with the peak at 1791.80 *m*/*z* including part of the determined region. These peaks were between those always detected (see Tables 2 and 3, 60 occurrences each). The peaks at 869.37 *m*/*z* and 658.81 *m*/*z* were also always detected, but their intensities changed in a specific manner, showing a clear tendency to decrease for the former and to increase for the latter even to the point of becoming the base peak of the mass spectra for checked day 88 and for both storage conditions.

Finally, the peptides (*m*/*z*) with five occurrences in the last day were tracked over time considering just their occurrences (bearing in mind that the analyses were always performed using the most representative peptides, which means that they were among the most intense and with a high level of reproducibility between replicates). Again, results for both storage con-

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

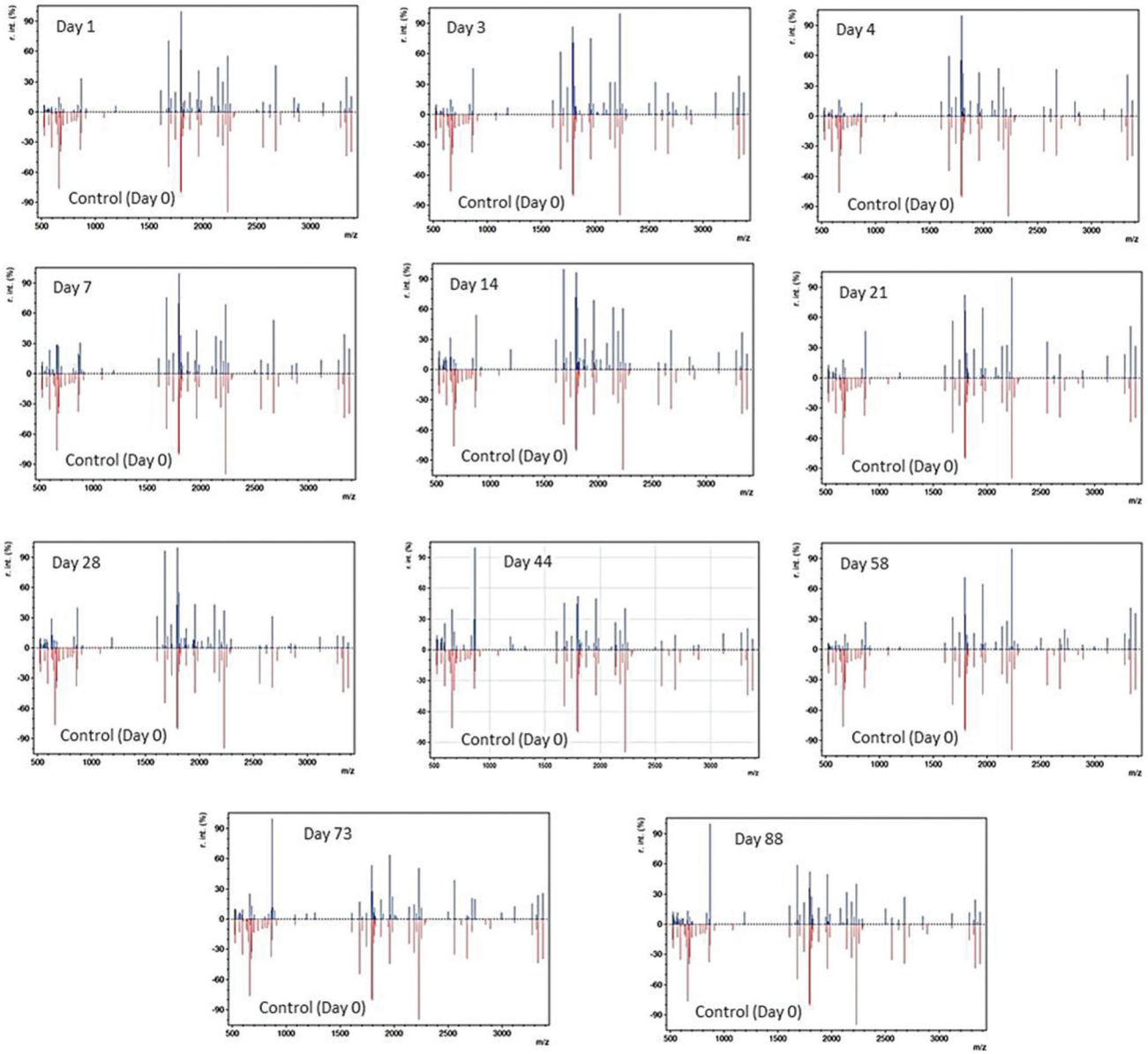

Fig. 6 Mass spectra of the most representative peptide over time *versus* control day. Frozen samples.

ditions were exactly the same. There were 9 peptides that were not detected on the control day including the most representative, *i.e.* 532.06 *m*/*z* (detected from day 3 onwards), 534.11 *m*/*z* (detected from day 3 onwards), 569.41 *m*/*z* (detected from day 58 onwards), 571.34 *m*/*z* (detected from day 1 onwards), 573.33 *m*/*z* (detected from day 3 onwards), 589.08 *m*/*z* (detected from day 14 onwards), 613.41 *m*/*z* (detected from day 58 onwards), 824.46 *m*/*z* (detected only on day 3 and the last day, day 88) and 2142.09 *m*/*z* (detected from day 1 onwards). A further four peptides were detected just once on day 0, although their occurrences increased later, *i.e.* 558.07 *m*/*z*, 835.45 *m*/*z*, 1682.22 *m*/*z* and 1964.89 *m*/*z*.

Therefore, gathering all the results obtained by the different ways of analysing the PMF of the most representative peptides, it is suggested that there were no differences regarding RTX preservation between the two storage conditions studied with the main modification on the RTX PMF happening after 24 hours of conservation. The results suggest that modifications occur from the first day, with a number of peptides not detected from the first checked day (day 1) and visible changes in the mass spectra of the most representative peptides of the last day compared to those of the control day.

Conclusion

The algorithm presented in this research is for clustering mass peaks (*m*/*z*) in order to enable comparison of complex mass spectral data with reliable matching. This algorithm can be

 This journal is © The Royal Society of Chemistry 2015

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

used with applications that involve complex data with unlabelled features, such as the PMF from the MALDI-TOF-MS of the RTX studied here in which the peaks are unlabelled. The algorithm takes into account the proportional error of the mass spectrometer equipment recording the m/z signals, and clusters the masses together according to their closeness and the maximum error accepted for each mass measurement, as indicated by the equipment specifications. The running time of the algorithm is almost negligible, thus enabling the user to analyse huge amounts of data within a few seconds. In the future, we plan to release a web-based implementation that will be available to the community.

We have successfully demonstrated the application of the algorithm for evaluating the chemical stability of a marketed mAb in a long term study covering the practical use of the anti-cancer drug, always bearing in mind that several analytical tools or methodologies have to be used to take a final decision on the stability of any mAbs. The algorithm was based on the PMF obtained by MALDI-TOF-MS. We also proposed the use of PMF replicates in order to obtain the most representative peptides, those always detected among the 100 most intense peptides and to focus the study on these. In this way, by applying the algorithm to all the PMFs obtained in the long term study, we brought to light particular aspects that are difficult to observe by visual comparison. Thanks to this approach, all the peptides in the PMF could be tracked between samples using different strategies, such as tracking over time the most representative peptides from the control day (in order to detect those that disappeared), tracking over time the most representative ones on the last day (to detect new peptides), and graphically comparing PMF for the control day *versus* each of the other checked days to evaluate changes in the PMF pattern (changes in normalized intensities).

In our particular long term study of RTX samples prepared under the usual hospital conditions, *i.e.* 1.0 mg mL^{-1} in NaCl 0.9%, and stored also under the usual hospital conditions refrigerated at 4 °C or frozen at −20 °C, we were surprised to find that the results clearly suggested that there were no differences between the two storage conditions, that is, the state of preservation of the RTX samples was the same when stored refrigerated at 4 °C or frozen at −20 °C. Our results also suggested that the main changes in the RTX structure occurred within the first 24 hours of storage. It should be checked whether these chemical alterations are correlated with modifications in the RTF biological functionality. This represents the core of the new research that we plan to perform in the context of a wider project, which is also to look for the particular modification by using a high-resolution and accurate mass spectrometer.

Conflict of interest

The authors confirm that the contents of this article are not affected by any conflict of interest.

Acknowledgements

Financial support was provided by the Project FIS:PI10/00201 (National Research Program, Instituto Carlos III, Ministerio de Economía y Competitividad, Spain), TIN2011-27696-C02-01 from the Ministerio de Economía y Competitividad, and P11-TIC-8001 from the Gobierno de Andalucía (Spain), CEI2014-MPB521 from CEI-BioTic Granada, and FEDER funds. The first author acknowledges support from a FPU scholarship from the Spanish Ministry of Education. The authors would like to thank the Hospital Pharmacy Unit of the University Hospital of San Cecilio (Granada, Spain) for kindly supplying all the medicine samples and the Biomedical Research Foundation "Alejandro Otero" (FIBAO) for the support given during the course of this research.

References

1 N. Sommerer, D. Centeno and M. Rossignol, Peptide Mass Fingerprinting, in *Plant Proteomics, Methods and Protocols*, ed. H. Thiellement, M. Zivy, C. Damerval and V. Méchin, Humana Press Inc., New Jersey, 2007, pp. 222–234.
2 B. Pramanik, G. Chen and M. Gross, *Protein and Peptide Mass Spectrometry in Drug Discovery*, John Wiley & Sons, New York, 2011.
3 N. D. Padliy and T. D. Wood, *Anal. Chim. Acta*, 2008, **627**, 162–168.
4 J. S. Morris, K. A. Baggerly, H. B. Gutstein and K. R. Coombes, *Methods Mol. Biol.*, 2010, **641**, 143–166.
5 S. Gibb and K. Strimmer, *Bioinformatics*, 2012, **28**, 2270–2271.
6 C. Yang, Z. He and W. Yu, *BMC Bioinf.*, 2009, **10**, 1–13.
7 P. Radivojac and O. Vitek (editors), *BMC Bioinf.*, 2012, **13**(supplement 16).
8 S. E. Reichenbach, X. Tian, A. A. Boateng, C. A. Mullen, C. Cordero and Q. Tao, *Anal. Chem.*, 2013, **85**, 4974–4981.
9 B. Wang, A. Fang, J. Heim, B. Bogdanov, S. Pugh, M. Libardoni and X. Zhang, *Anal. Chem.*, 2010, **82**, 5069–5081.
10 Q. P. He, J. Wang, J. A. Mobley, J. Richman and W. Grizzle, *Cancer Inf.*, 2011, **10**, 65–82.
11 J. R. Arnold and J. P. Reilly, *Rapid Commun. Mass Spectrom.*, 1998, **12**, 630–636.
12 M. Strohalm, D. Kavan, P. Novák, M. Volný and V. Havlícek, *Anal. Chem.*, 2010, **82**, 4648–4651.
13 Open Source Tools for Mass Spectrometry Analysis: http://strimmerlab.org/notes/mass-spectrometry.html (accessed October 2014).
14 S. Kim, A. Fang, B. Wang, J. Jeong and X. Zhang, *Bioinformatics*, 2011, **27**, 1660–1666.
15 X. Zhang, A. Fang, C. P. Riley, M. Wang, F. E. Regnier and C. Buck, *Anal. Chim. Acta*, 2010, **664**, 101–113.
16 Scientific discussion Rituximab (2005) EMEA: http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_

Published on 24 December 2014. Downloaded by Universidad de Granada on 13/03/2015 00:20:38.

Scientific_Discussion/human/000165/WC500025817.pdf. Accessed 22 July 2014 (accessed October 2014).
17 M. C. Cheung, A. E. Haynes, R. M. Meyer, A. Stevens and K. R. Imrie, *Cancer Treat. Rev.*, 2007, **33**, 161–176.
18 J. C. Edwards, L. Szczepanski, J. Szechinski, A. Filipowicz-Sosnowska, P. Emery, D. R. Close, R. M. Stevens and T. Shaw, *N. Engl. J. Med.*, 2004, **350**, 2572–2581.
19 S. Rosati, N. J. Thompson and A. J. R. Heck, *TrAC, Trends Anal. Chem.*, 2013, **48**, 72–80.
20 C. Bardina, A. Astier, A. Vulto, G. Sewell, J. Vignerone, R. Trittler, M. Daoupharsg, M. Paul, M. Trojniak and F. Pingueti, *Ann. Pharm. Fr.*, 2011, **69**, 221–231.
21 M. Paul, V. Vieillard, E. Jaccoulet and A. Astier, *Int. J. Pharm.*, 2012, **436**(2012), 282–290.
22 A. Antoniadis, J. Bigot and S. Lambert-Lacroix, *J. Soc. Fr. Stat.*, 2010, **151**, 17–37.
23 D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, Cambridge, 2012.
24 N. Memarsadeghi, N. S. Netanyahu and J. LeMoigne, *Int. J. Comp. Geometry Appl.*, 2007, **17**, 71–103.
25 M. Ester, H. P. Kriegel and X. Xu, *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96) Portland*, Oregon, 1996.
26 EMEA product information 25/05/2012 MabThera -EMEA/H/C/000165 -II/0077GAnnex I: Summary of product characteristics: http://www.emea.europa.eu/docs/en_GB/document_library/EPAR_Product_Information/human/000165/WC500025821.pdf Accessed 25 July 2014, (accessed October 2014).
27 T. H. J. Niedermeyer and M. Strohalm, *PLoS One*, 2012, **7**, e44913.
28 M. Strohalm, M. Hassman, B. Košata and M. Kodíček, *Rapid Commun. Mass Spectrom.*, 2008, **22**, 905–908.
29 D. Nebija, H. Kopelent-Frank, E. Urban, C. R. Noe and B. Lachmann, *J. Pharm. Biomed. Anal.*, 2011, **56**, 684–691.
30 M. E. Reff, K. Carner, K. S. Chambers, P. C. Chinn, J. E. Leonard, R. Raab, R. A. Newman, N. Hanna and D. R. Anderson, *Blood*, 1994, **83**, 435–445.
31 J. Du, H. Wang, C. Zhong, B. Peng, M. Zhang, B. Li, S. Hou, Y. Guo and J. Ding, *Mol. Immunol.*, 2008, **45**, 2861–2868.

 This journal is © The Royal Society of Chemistry 2015