



The author identified by his method: EuPA YPIC challenge solved

M.I. Indeykina^{a,b,c,*}, D.A. Podgrudkov^d, A.S. Kononikhin^{b,c,e,*}

^a Emanuel Institute of Biochemical Physics of the Russian Academy of Sciences, Moscow, Russia

^b Moscow Institute of Physics and Technology (State University), Moscow, Russia

^c Talrose Institute for Energy problems of Chemical Physics of the Russian Academy of Sciences, Moscow, Russia

^d Lomonosov Moscow State University, Physics Department, Moscow, Russia

^e V.I. Kulakov National Medical Research Center of Obstetrics, Gynecology and Perinatology, Moscow, Russia



ARTICLE INFO

Keywords:

EuPA YPIC challenge
De novo sequencing
Mass-spectrometry

ABSTRACT

Here we present the results of our attempt on the EuPA YPIC challenge. The task was to sequence the provided synthetic peptides, build the sentence encrypted in them and determine from which book the sentence originated.

The task itself, while holding no direct scientific value, offers an insight in less formal terms (for participants at least) on how the overall process of a scientific study of a “new protein” looks like. Hence, we decided to look at the challenge as if it was a general task of sequencing an unknown protein from an unusual proteome database. To solve the task we used LC–MS/MS, MALDI-MS and de novo sequencing. A combination of two MS instruments and de novo MS/MS data analysis make it possible to sequence new peptides and proteins not yet present in proteomic databases.

1. Introduction

The EuPA YPIC challenge (for full rules and conditions see [1]) set an interesting task of identifying a book by a quote encoded in 19 peptide sequences, each of which contained 1–5 words with spaces and punctuation removed. Letters that are not used as one-letter codes of amino acids, but were required for the quote were encoded using specific post-translation modifications. The task was to sequence the provided peptides, build the sentence and determine from which book the sentence originated.

The task itself, while holding no direct scientific value, offers an insight in less formal terms (for participants at least) on how the overall process of a scientific study of a “new protein” looks like. Hence we decided to look at the challenge as if it was a general task of sequencing an unknown protein from an unusual proteomic database.

Usually for such tasks (antibody sequencing for example) the isolated protein is digested by several enzymes of different specificity and subjected to LC MS/MS analysis. The derived peptides are separated on LC, followed by accurate mass measurement and fragmentation of each peptide. The derived MS/MS data is subjected to de novo sequencing algorithms in which the sequence of the parent ion is determined from the differences in masses of the derived fragments according to the rules of peptide dissociation. After sequencing of the peptides the sequence of the whole protein is reconstructed by analyzing the intersections of

peptides obtained under the action of enzymes with different amino acid residue specificities.

In the presented study the “protein” was already divided into peptides, so the digestion step was omitted, and since there were no peptide intersections to analyze a slightly unusual database approach was used to identify the order of the peptides and reconstruct the whole “protein” sequence.

2. Materials and methods

The sample vial arrived in the lab in the middle of July 2017. As was stated in the description it should have contained 40 µl of the peptide mixture sample containing roughly 0.5 nmol of each peptide in 30% ACN. Unfortunately, the contents have dried and were resuspended in 40 µl 30% ACN (Merck). 1 µl of the resuspended sample was taken and diluted with 20 µl of HPLC grade H₂O (Merck). For a quick first look at the sample and to check for sufficiency of the peptide concentration levels and sample/solvent quality, MALDI TOF MS spectra were obtained on a Bruker microflex instrument using the HCCA matrix. Further LC MS/MS experiments were carried out on an Agilent 1100 nanoHPLC system coupled to a 7 T Thermo Finnigan LTQ FT Ultra mass-spectrometer with a nanoESI source. The peptide mixture was separated on a homemade C18 capillary column (i.d. 75 µm × length 12 cm, Reprosil-Pur Basic C18, 3 µm, 100 Å; Dr. Maisch HPLC GmbH,

* Corresponding authors at: Moscow Institute of Physics and Technology (State University), Moscow, Russia.

E-mail addresses: mariind@yandex.ru (M.I. Indeykina), alex.kononikhin@gmail.com (A.S. Kononikhin).

<https://doi.org/10.1016/j.euprot.2018.10.001>

Received 9 September 2017; Received in revised form 8 October 2018; Accepted 9 October 2018

Available online 13 October 2018

2212-9685/ © 2018 Published by Elsevier B.V. on behalf of European Proteomics Association (EuPA). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Germany). A 140 min total separation at a flow rate of 0.3 $\mu\text{L}/\text{min}$ (solution A – 0.1% formic acid in H_2O , solution B – 100% ACN) with the following gradient set up was used:

- 0–15 min: 3% buffer B
- 15–85 min: linear gradient from 3–50% of buffer B
- 85–105 min: linear gradient from 50–90% of buffer B
- 105–115 min: 90% of buffer B
- 115–125 min: linear gradient from 90–4% of buffer B
- 125–140 min: re-equilibration of the column in 3% buffer B

The MS settings were as follows:

1. Masses of parent ions and their charge states were measured in the ICR cell with high mass accuracy in the m/z range of 200–2000 with a resolution of 50,000 at m/z 400 at AGC Target setting of 1e6, and maximal injection time of 500 ms.

2. Five most intense ions with a detected charge state in each MS scan were subjected to MS/MS fragmentation in a data-dependent mode in the linear ion trap

- a Dynamic exclusion was used to prevent re-acquisition of MS/MS spectra of previously selected ions – after acquisition of 2 fragmentation spectra over a 15 s time window the parent mass with a tolerance of ± 2 ppm was added to an exclusion list (holding maximum 300 entries) and thus excluded from further fragmentation selection for the next 30 s.
- b The m/z range in MS/MS mode was determined from the precursor ion m/z
- c Collision-induced dissociation (CID) was used for fragmenting the parent ions in a window with a width of 2 and applied energy of 25%.
- d For the fragment spectra acquisition the AGC Target setting was 1e4 and maximal injection time was 150 ms.

10 full LC MS/MS runs with injection of 2 μL of 50%ACN as sample were carried out to wash the system prior to analysis followed by 5 experimental runs with 1 μL of the peptide sample (Fig. 1).

The resulting raw files were uploaded into the PEAKS Studio v.8.0. (Bioinformatics Solutions Inc) software package.

A de novo search with the following parameters was performed (Fig. 2).

- Parent Mass Error Tolerance: 15.0 ppm
- Fragment Mass Error Tolerance: 0.5 Da
- Enzyme: None
- Variable Modifications:
- Acetylation: 42.01
- Artif: 89.97
- Methylation (R): 14.02
- Phosphorylation: 79.97

Besides the modifications indicated in the technical documentation for the sample a set of usually occurring during sample preparation artifact modifications was added, such as

- Deamidation (NQ): 0.98
- Oxidation (HW): 15.99
- Oxidation (M): 15.99
- Sodium adduct: 21.98
- Max Variable PTM Per Peptide: 10
- Report # Peptides: 5

After de novo an identification search over the SwissProt database was run to filter out the contaminants remaining in the system from previous routine analyses with a standard set of parameters:

- Parent Mass Error Tolerance: 15.0 ppm
- Fragment Mass Error Tolerance: 0.5 Da
- Precursor Mass Search Type: monoisotopic
- Enzyme: None
- Max Missed Cleavages: 100
- Non-specific Cleavage: both
- Variable Modifications:

- Carbamidomethylation: 57.02
- Deamidation (NQ): 0.98
- Oxidation (HW): 15.99
- Oxidation (M): 15.99
- Sodium adduct: 21.98
- Phosphorylation (STY): 79.97
- Max Variable PTM per Peptide: 3
- Database: Swissprot_human

After filtering out all identified sequences belonging to contaminant proteins left in the LCMS system the results were sorted by RT, sequence and mass, and looked through manually for words.

Although, already at this stage a number of words and peptides were decoded, the need for some automated method to cut down the number of repeating possible sequences and weighting of the results became evident and was realized using the Mathworks MATLAB software package.

To further reduce the number of sequences, the de novo output set was checked for duplicates using the following algorithm (Fig. 3). For this, each sequence one by one was removed from the set, matched against the list of already checked sequences (the very first sequence wasn't matched but simply moved to that list) and, if already present, discarded from further consideration and if not, added to that list. Then it was checked against the remaining sequences in the original set. If found, the local confidence (l_i) for each amino-acid and average local confidence for the amino-acids were updated to reflect this. The new local confidence was changed to either the highest l_i found or $1 - (1 - l_i) / \text{number of duplicates}$ (l_i - average local confidence), whichever was higher. After that the average local confidence was recalculated as:

$$\bar{P} = \sqrt[n]{\prod_{i=1}^N l_i} \quad (1)$$

where N is the length of the sequence.

Then an algorithm aimed at identifying sequences that most probably originated from the same peptide was implemented. For each pair of sequences their every possible overlapping combination was checked using the local confidence information, e.g. for each overlap the elements (i -th of one and j -th of the other) of the sequences were compared pair by pair. Only pairs in which both l_i and l_j were higher than 50% were used. If no such pair was found for the specific overlap, then its probability was estimated as 0. For other pairs the total probability of the overlap was estimated as

$$P = \prod_k p_k$$

where $p_k = l_i l_j$ if the pair consisted of two equal amino-acids and $p_k = (1 - l_i) l_j + l_i (1 - l_j)$ if the pair consisted of different amino-acids.

Since the lengths of the overlapping parts of the sequences were different, the average confidence $\bar{P} = \sqrt[n]{P}$ (where n is the number of pairs used in comparison) was used to find the most probable overlap for each pair of sequences (Fig. 4).

If the calculated probability was high (above 0.65), the resulting new sequence was built and stored with its own local confidence and average local confidence info. The new sequence consisted of amino-acids from the non-overlapping parts from two original sequences and the amino-acids with highest local confidence of the respective pair in the overlapping part. The local confidence for the new sequence was built the same way. The average local confidence was estimated again using Eq. (1).

After this the duplicate removal procedure was repeated using the same algorithm as described above to cut down the newly obtained repeats.

3. Results and discussion

Despite the multiple preparative wash runs of the system the reconstructed by de novo sequences (around 4500) were first filtered to

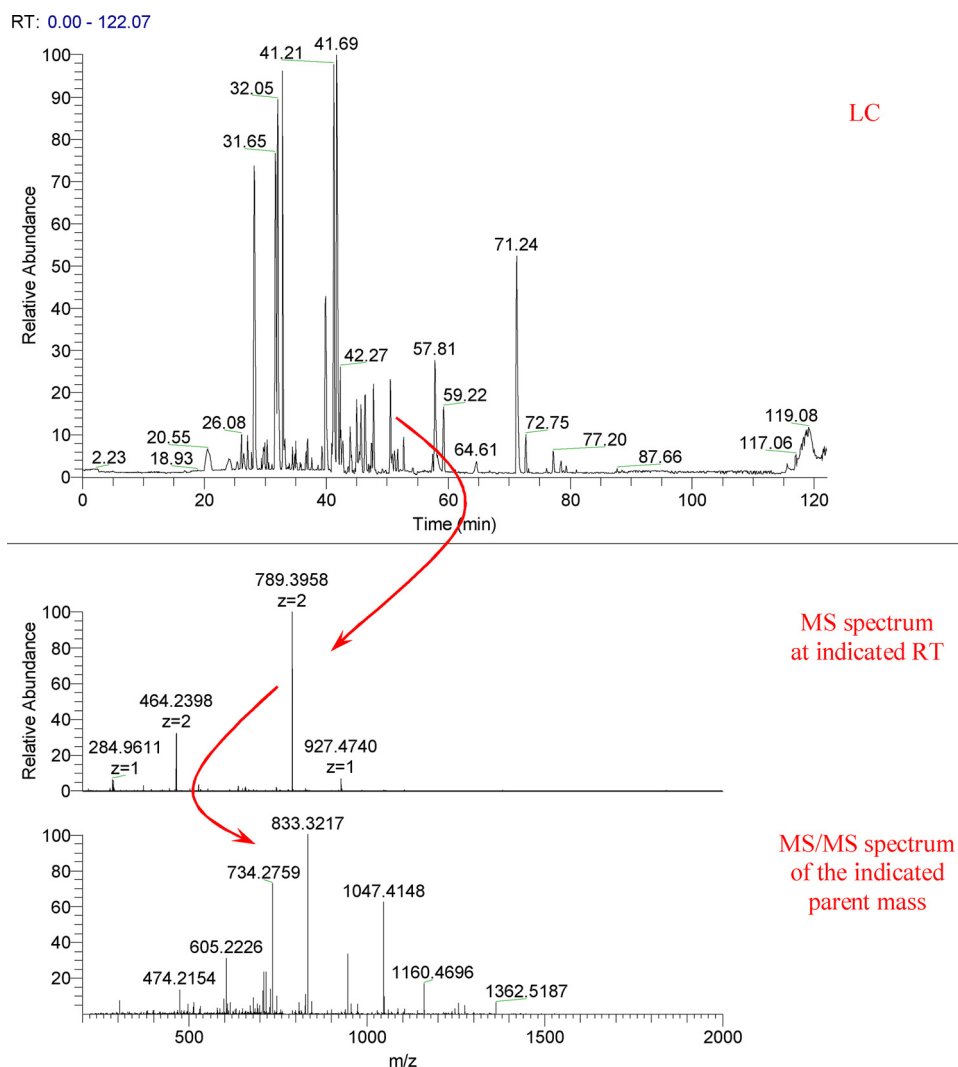


Fig. 1. An example of LC MS/MS data. At every retention time 5 most intense ions from the parent mass-spectrum acquired at high resolution and mass-accuracy in the ICR cell were isolated and fragmented using CID with the detection of fragment ions with high sensitivity in the ion trap.

remove the remaining contaminants from previous runs – usual non-artificial peptide sequences identified by the SwissProt-human database (about 250). The remaining de novo results were sorted by RT, sequence and mass and looked through manually for words, with primary attention to those containing the unusual modifications, encoding the missing letters, since their source of origin was of no doubt (Fig. 2).

After screening of about 1000 of the most high scoring de novo sequences, about 20 words and at least 4 peptides were identified, but the need for optimization, i.e. approaches to cut down the number of repeating possible sequences, and automatization became obvious.

First, duplicates were removed from the set of sequences and the confidence levels of the sequences were updated to represent the higher reliability of those with multiple repeats. In the result of this step about 1000 duplicates were removed.

The remaining set was subjected to a pair matching process. At low confidence levels this procedure generated numerous random sequences, but at high average confidence level threshold (0.90) it effectively identified the sequences originating from the same peptide. The new list was then again checked for duplicates. At this point several sequences (total of 65 and most of them were spotted earlier during different stages) were identified as being reconstructed correctly.

ANALYLSREQ - analysis req(uires/d?)

THEMETHODLS - the method is

ANYOTHERMETHOD - any other method

ANDDOESNOTREQ - and does not req(uires/d?)

ENSLTLVEMORE - sensitive more?

SOEVENTHATOF - so even that of

WLTHFAR - with far

THLSTHAN - this than

LFEELSU - I feel so?/su(re)

SPECGTRUM - spectrum?

NTOFMATERLAL - (amou)nt of material

In case of a usual proteomic study the de novo and MS/MS results are searched against various proteome bases, such as SwissProt [2] for example, to identify the source protein or even organism, using MS specialized “search engines”, such as MASCOT, Comet, Xtandem! and others. For this case another search engine and database were used — Google and the Google Books collection [3]. The query formed from the identified sequences (“any other method” so even that of “and does not require” “analysis requires” “sensitive more”) returned a few books¹ with the phrase “I feel sure that there are many problems in chemistry which could be solved with far greater ease by this than by any other method. The method is surprisingly sensitive — more so even than that of Spectrum Analysis, requires, an infinitesimal amount of material, and

¹ We randomly picked the *Handbook on the Physics and Chemistry of Rare Earths: Non-Metallic Compounds* by K. Gscheidner and L. Eyring (1979, p.359), but all of the books contained the same quote.

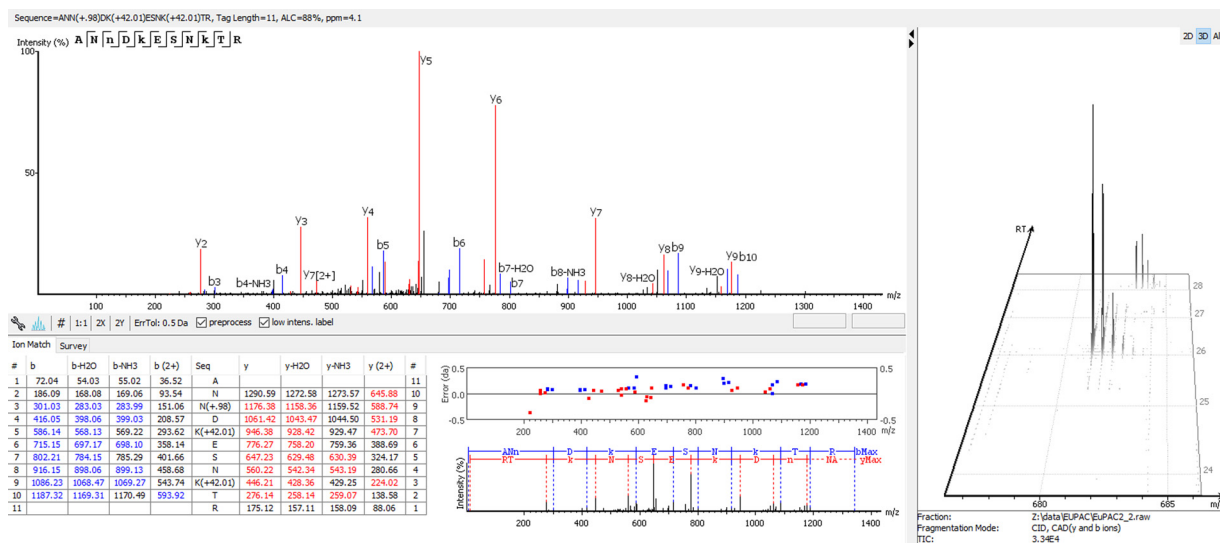


Fig. 2. Example of peptide sequencing results basing on MS/MS data. Part of the MS1 spectrum showing the parent ion (right 3D spectrum), mass spectrum of observed fragment ions (top left 2D spectrum), table of theoretical masses for the different fragment ion series (bottom left table), plot of mass errors between observed and theoretical values for detected fragment ions (central middle panel), peptide sequencing results using the b and y ion series (bottom middle panel). The observed b and y fragment ions are shown in blue and red respectively. The mass difference between two consecutive ions of the same type is equal to the mass of the corresponding amino acid residue, thus allowing its identification. Analysis of a consecutive series of ions of each type allows to reconstruct the sequence of the residues in the peptide. Unobserved or unidentified fragment ions are shown in black.

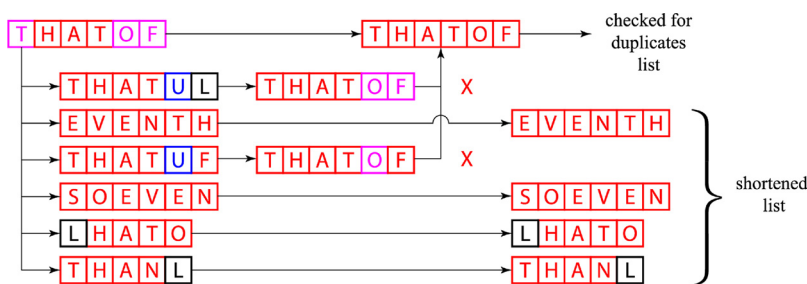


Fig. 3. Principal schema of the duplicate comparison algorithm. Each sequence is matched to the rest remaining in the original set in search of duplicates taking into account the confidence of identification levels for each amino acid. In case of coincidences the confidence in correct identification is increased and duplicating sequences are removed shortening the list which is returned to the beginning of the algorithm cycle for comparison with the next sequence. The program's confidence of the correct identification of each amino acid is reflected by its color: red being most confident, followed by magenta, blue, and black represents least confident residues.

does not require this to be specially purified: the technique is not difficult if appliances for high vacua are available". The phrase was used as a quote and now having it in full text a quick search in Google returned its origin: preface of the *Rays of Positive Electricity and Their Application to Chemical Analysis* [4] by J.J. Thompson.

Additional confirmation of the correctness of identification was that there are 6 letters of the English alphabet that do not encode amino-acids: B, J, O, U, X and Z. Three of these letters were encoded for this challenge by specific PTMs (B, O and U), leaving 3 more out of consideration. Though these letters (J, X and Z) are among the least used in the modern English language, they should not be present in the original quote.

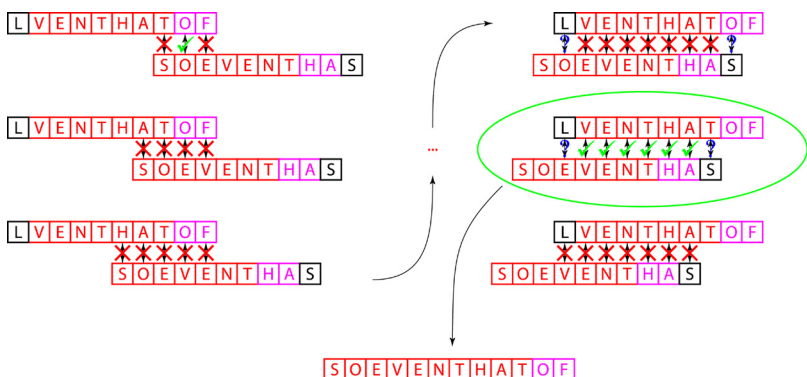


Fig. 4. Sequence overlap analysis algorithm. For each pair of sequences their every possible overlapping combination was checked using the local confidence information, e.g. for each overlap the elements (i-th of one and j-th of the other) of the sequences were compared pair by pair. The most probable overlap for each pair of sequences was calculated and if it was high, the resulting new sequence was built and stored. The new sequence consisted of amino-acids from the non-overlapping parts from two original sequences and the amino-acids with highest local confidence of the respective pair in the overlapping part.

Thomson [[5]]; from F.W. Aston [5, p.661 as an example of an author close in time and area of research; from R. Feynmann [6] as an example of modern scientific physics; the PEAKS Studio introduction article [7], a random programming textbook, a proteomics review article [8], and Watson and Crick [9] as examples of modern and contemporary scientific English from other areas of sciences; and J.J.R. Tolkien [10] and W. Shakespeare [11] as standards of literary English. These extra quotes serve as an additional test against the pileup of low confident falsely identified sequences from the background noise that can theoretically contribute to the coverage of a peptide while not being actually present or belonging to it.

The following coverage was observed for the selected quotes (at FDR of 1%):

Original – 66% (see Fig. 4 and discussion below)

Proteomics – 9% (Analysis & (amou)nts of material)

PEAKS article – 5% (Sensitiv(ity), (ident)ified, require)

There were no hits on the poetic or other selected quotes. Of those quotes that had non-zero coverage it was on words present in the original phrase, such as “analysis”, “material”, “sensitive” and “require” and no new words were sequenced.

The original quote did not show full coverage (Fig. 5) – the end of the sentence was completely missing from our data and probably from the challenge in whole, since this phrase is often quoted only to this place and instead of a colon a point is placed. Also, in the challenge set up it was stated that two words should be missing and the meaning of one of them should become obvious in combination with “small amount”. Since the absence of the second word or its meaning is not explained by the challenge designers, we supposed that it is an article and these missing words are thus “an infinitesimal”, which remain totally non-covered in our analysis. Considering these absences, the observed coverage reaches 90%.

Additionally, several peculiarities were observed: the original phrase in the middle holds “SO EVEN THAN THAT OF”, but the peptide in the mixture is “SO EVEN THAT OF” with “THAN” omitted. Also, both words starting with SPEC (specially and spectrum) were de novo sequenced with the stated set of modifications as SPECG or SPEGC, and, when a PTM search for not indicated modifications was run, were reconstructed as carbamidomethylated at the C residue, though no such procedure was done in our sample preparation and is not mentioned in the initial sample description. As was found further, this modification is

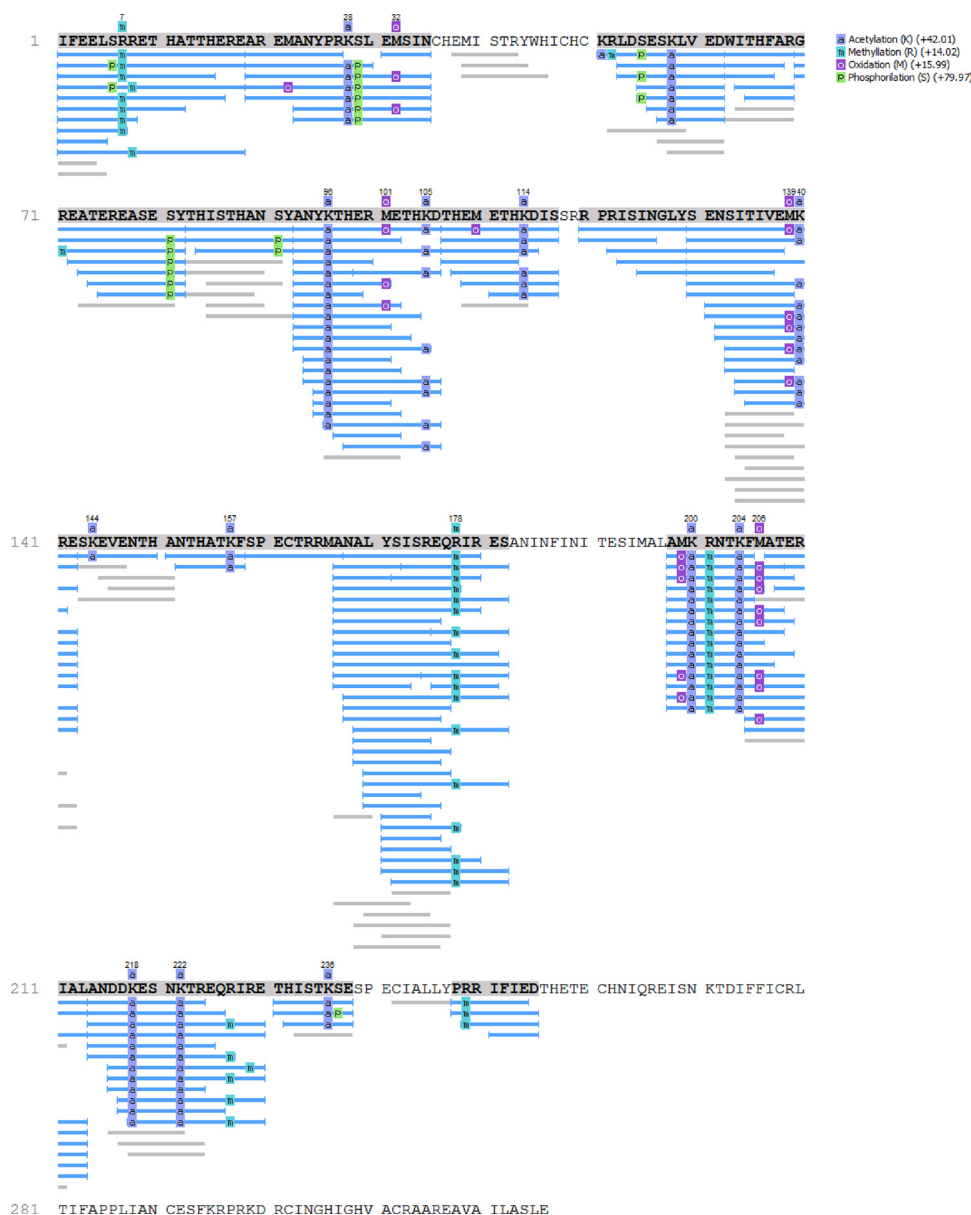


Fig. 5. Sequence coverage of the original quote with only modifications from the sample description allowed.

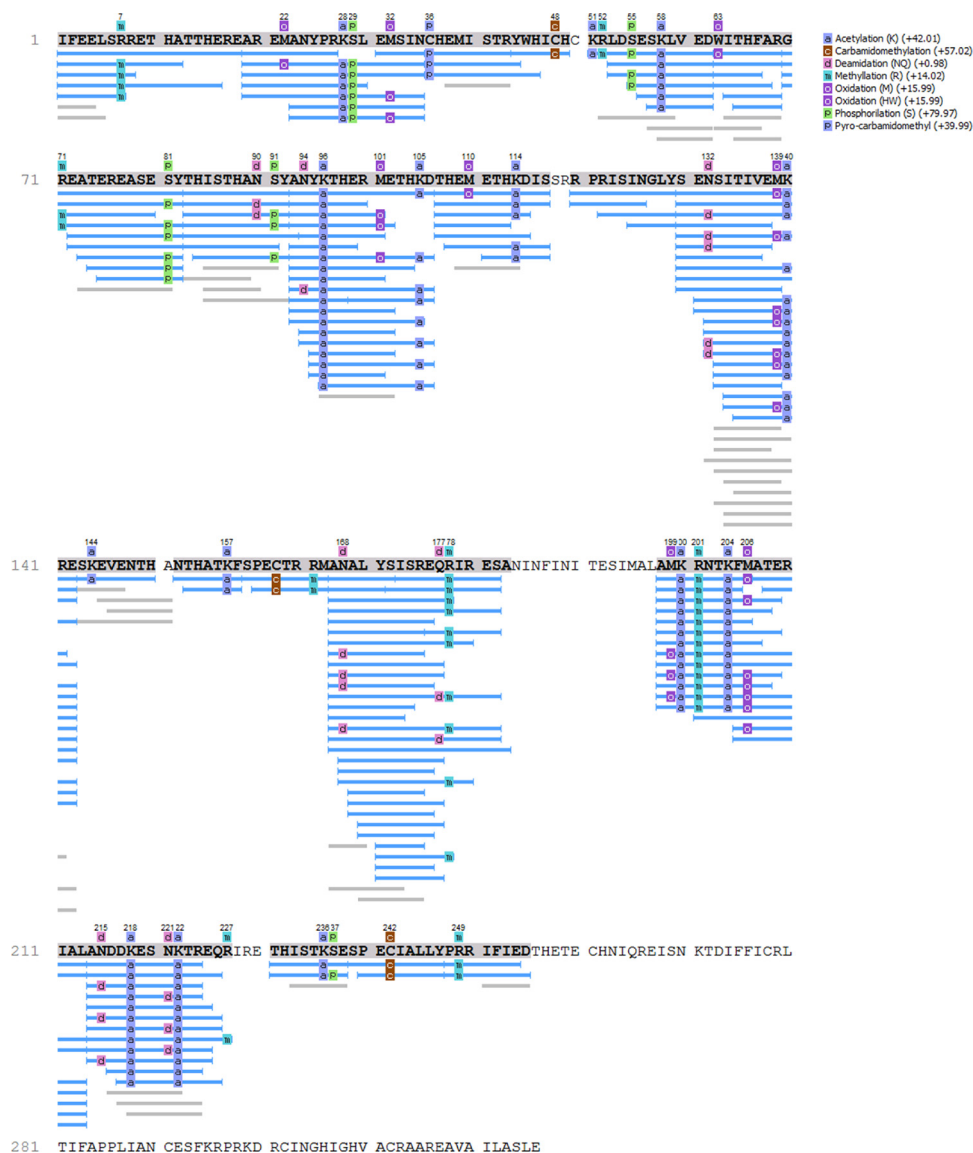


Fig. 6. Quote sequence coverage when additional identified modifications are allowed.

also registered for other sequence fragments containing the C residue (“which”, “could”) and several contaminant peptides as well. Also the word “chemistry” was sequenced as starting with EE or sodiated EE, with a mass difference of 39.99, and was identified by the additional PTM search as carrying the pyro-carbamidomethyl modification (with the mass of +39.99). Thus with the addition of these two modifications and usual artifacts, such as oxidation and deamidation, to the list an almost full coverage was obtained (Fig. 6).

After obtaining the sequence coverage, the quote was divided into separate peptides basing on the evident connection points to double check their number, presence in MALDI and LC MS spectra and absence of unidentified items (Figs. 7 and 8). For reliability MALDI spectra were reacquired on a BRUKER Ultraflex instrument, which allows better resolution and mass accuracy than its smaller microflex brother, used for the first MALDI screening. Peptides “could be solved”, “spectrum” and “specially” were observed by all instruments only in their carbamido-methylated forms. Peptide “this to be” was detected only by LCMS and absent from all MALDI spectra, and long and heavy peptides such as “I feel sure that there”, “are many problems in”, “amount of material” and “and does not require” are absent or present at very low intensities in their full size in LC MS, and are presented in these spectra by their fragments of various lengths.

Besides the expected masses of peptides and their various fragments, several high intensity peaks were present in both MALDI and LC MS spectra and were not identified. The most standing out in the MALDI spectra of such peptides (m/z 1739.4) was found basing on its mass in the de novo results sequenced as M(+15.99) PC(+57.02)TEDYLSLLNR and subjected to a BLAST search, which returned a very close identity to BSA with a difference in I/L, which are indistinguishable by standard MS approaches. This sequence differs from the human albumin and so was not identified during our contaminant search, which was oriented on excluding contaminants coming from the instrument, since no work was done in the lab on other species for a significant period of time and thus contaminants from other species were not expected. Since these peptides were present on all 3 instruments, in various spots of both MALDI targets and both sample vials, we supposed that these contaminants originated from the sample itself and a sample contaminant search was performed using the full SwissProt database. This allowed to identify all remaining high intensity peaks in all spectra, thus closing the need for further validation. It should also be noted that these peptides were also carbamido-methylated.

In the rules it was mentioned that in one of the words a protection group (+89.97) probably on serine remained attached, but we were not

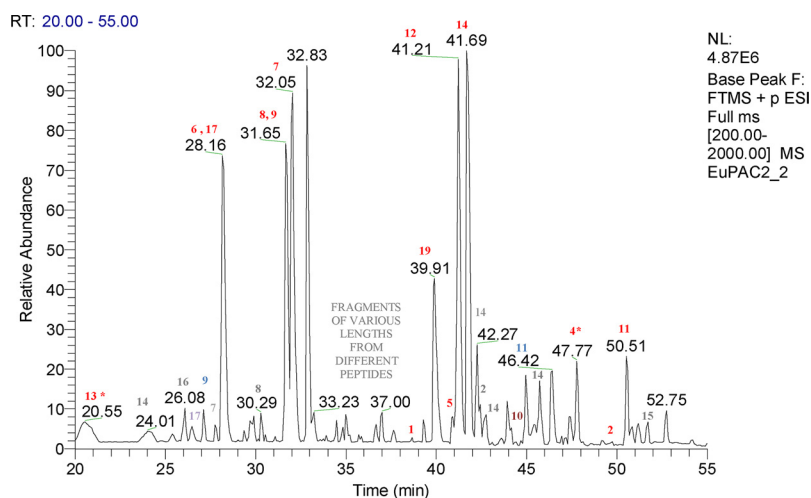


Fig. 7. Annotation of LC peaks to peptides – * – carbamidomethylated, red – full peptide, grey – part of peptide, blue - oxidized peptide. Numbers correspond to the order of the peptides in the quote sequence.

able to observe this modification, probably due to insufficient accuracy of the provided mass value, since changing the specificity of this modification from S to X also gave no result. If it is actually localized on S, then from our LC MS data the only place where it may be present is on the first S of the word “SURPRISINGLY”, which for some reason totally lacks coverage, while the neighboring residues are easily observed. But in the MALDI spectra a peak corresponding to this peptide without this modification is clearly present (10 – indicated by a darker red color) and cannot be explained by contaminants or fragments of other peptides. If the proposition on amino acid specificity provided by the organizers is not correct, then it may be on the first C of the word “could”, since it also lacks coverage, but though the software for some reason does not show coverage, the corresponding peaks without such modification seem to be present in the sample spectra. Therefore, the question on the presence and localization of this modification remains open.

4. Conclusions

As Sir Thomson believed the capabilities of mass-spectrometry for unraveling the unknown especially using modern instruments and in combination with modern data analysis software are extremely high. Using de novo MS/MS data analysis approaches it is possible to sequence new peptides and proteins not yet present in proteomic databases. But thoughtful and careful validation and consideration of contaminants, PTMs and possible sequencing/synthesis errors is important to obtain reliable results.

Conflict of interest

The authors declare no competing financial interest.

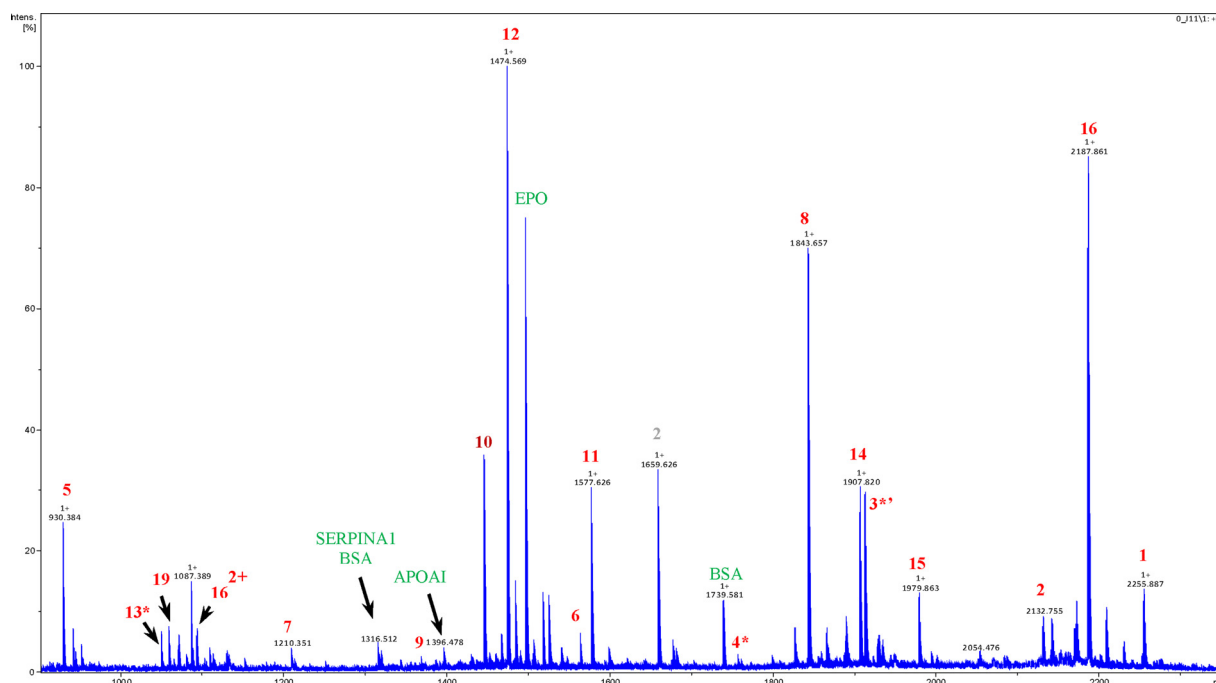


Fig. 8. Annotation of MALDI peaks to peptides – * – carbamidomethylated, red – full peptide, grey – part of peptide, green – contaminants. Numbers correspond to the order of the peptides in the quote sequence.

Acknowledgements

Authors wish to thank EuPA for providing such an interesting exercise, Katerina Poverennaya (IBMC RAS) for her organization efforts for the Russian EuPA members, Feodor Lukianov for his heroic effort on keeping the lab IT-infrastructure up and running, MIPT for access to the Ultraflex instrument and IBCP RAS center for collective use for providing the rest of equipment and resources. The study was supported by RFBR grants no. 16-54-21011_SNF_a, 17-08-01537 A.

References

- [1] <http://eupa.org/ypic/the-challenge/>.
- [2] www.uniprot.org.
- [3] Google Books // <http://books.google.com>.
- [4] J.J. Thompson, Rays of Positive Electricity and Their Application to Chemical Analysis, (1913).
- [5] F.W. Aston, The mass-spectra of chemical elements, Philos. Mag. 39 (1.233) (1920) 611–625 series 6.
- [6] R. Feynman, Surely You Are Joking, Mr. Feynman! Chapter “Cargo Cult Science”, (1985).
- [7] Jing Zhang, Lei Xin, Baozhen Shan et al., PEAKS DB: De Novo sequencing assisted database search for sensitive and accurate peptide identification, Mol. Cell. Proteom. 11 (2012).
- [8] M. Geisow, Proteomics: one small step for a digital computer, one giant leap for human kind, Nat. Biotechnol. 16 (1.2) (1998) 206.
- [9] J.D. Watson, F.H.C. Crick, A structure for deoxyribose nucleic acid //, Nature 171 (1953) 737–738.
- [10] J.R.R. Tolkien Lord of the rings, The Verse of the Rings.
- [11] W. Shakespeare, Henry V, Act III, first lines of “Cry Gog for Harry, England and Saint George”.