Manuscript Details

Manuscript number	EUPROT_2017_15
Title	The author identified by his method: EuPA YPIC challenge solved.
Short title	EuPA YPIC challenge solved
Article type	Full length article

Abstract

Here we present the results of our attempt on the EuPA YPIC challenge

Keywords	EuPA YPIC challenge, de novo sequencing, mass-spectrometry
Corresponding Author	Alexey Kononikhin
Corresponding Author's Institution	MIPT
Order of Authors	Maria Indeykina, Dmitriy Podgrudkov, Alexey Kononikhin
Suggested reviewers	Dmitry Avtonomov, Dmitry Galetskiy

Submission Files Included in this PDF

File Name [File Type]

Cover-letter.docx [Cover Letter]

Manuscript_EuPA YPIC_Indeykina-Kononikhin.docx [Manuscript File]

Conflict of interest.docx [Conflict of Interest]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Dear Editor,

Here we present the results of our attempt on the EuPA YPIC challenge. The task was to sequence the peptides, build the sentence and find out from which book that 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 the proteome looks like. Hence we decided to look at the challenge as if it was a general task of identifying a protein from an unusual proteome database. To solve the task we use LC-MS/MS, MALDI-MS and de novo sequencing. 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.

We wish to thank EuPA for providing such an interesting exercise, Katerina Poverennaya (IBMC RAS) for her organization efforts for Russians' EuPA members.

We will appreciate considering our paper for review!

With regards,

Authors

Title

The author identified by his method: EuPA YPIC challenge solved.

Authors

M.I. Indeykina*[1,2,3], D.A. Podgrudkov[4], A.S. Kononikhin*[1,2,3]

Affiliations

[1]Emanuel Institute of Biochemical Physics of the Russian Academy of Sciences

[2]Talrose Institute for Energy problems of Chemical Physics of the Russian Academy of Sciences

[3]Moscow Institute of Physics and Technology (State University)

[4]Lomonosov Moscow State University, Physics Department

Corresponding author:

M.I. Indeykina, mariind@yandex.ru

A.S. Kononikhin, alex.kononikhin@gmail.com

Abstract

Here we present the results of our attempt on the EuPA YPIC challenge. The task was to sequence the peptides, build the sentence and find out from which book that 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 the proteome looks like. Hence we decided to look at the challenge as if it was a general task of identifying a protein from an unusual proteome database. To solve the task we use LC-MS/MS, MALDI-MS and de novo sequencing. 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.

Key words

EuPA YPIC challenge, de novo sequencing, mass-spectrometry,

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 peptides, build the sentence and find out from which book that 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 the proteome looks like. Hence we decided to look at the challenge as if it was a general task of identifying a protein from an unusual proteome database.

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 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). 1µl of the resuspended sample was taken and diluted with 20 µl of HPLC grade H_2O (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 on a Bruker Ultraflex instrument were obtained using the HCCA matrix. Further LC MS/MS experiments were carried out on an Agilent 1100 nanoHPLC system coupled to a 7T 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, Germany).A 140 min total separation at a flow rate of 0.3 µL/min (solution A – 0.1% FA in H2O, solution B – 100% ACN) with the following gradient set up was used:

0–15 min: 3 % buffer B

15–85 min: linear gradient form 3–50 % of buffer B

85–105 min: linear gradient form 50–90 % of buffer B

105–115 min: 90 % of buffer B

115–125 min: linear gradient form 90–4 % of buffer B

125-140 min: reequilibration 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 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 15s 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 30s.

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 the window width of 2 and applied energy of 25%.

d. For the fragmentation spectra acquisition the AGC Target setting was 1e4 and maximal injection time 150 ms.

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

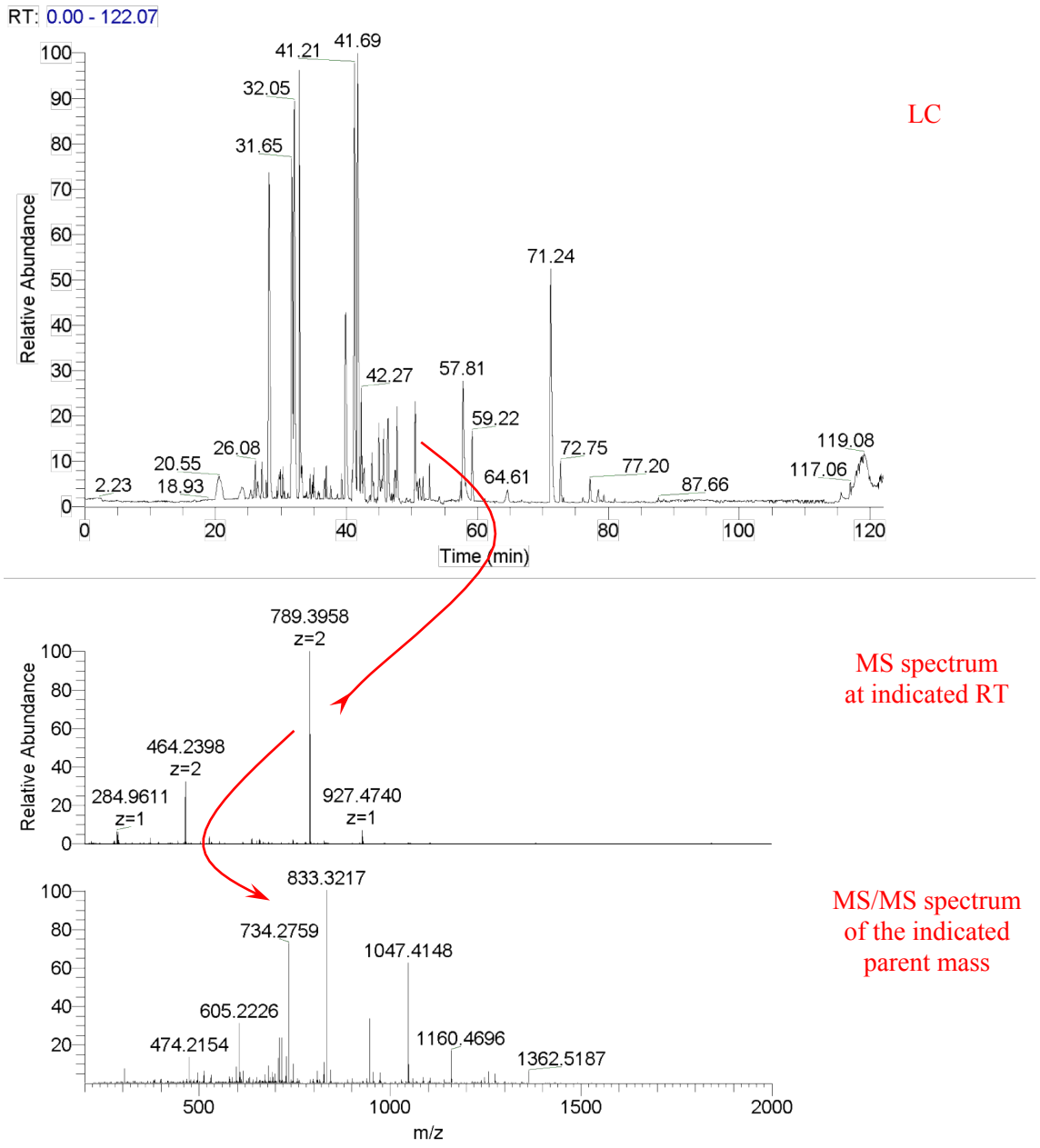


Figure 1. An example of LC MS/MS results.

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.

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 system the results were sorted by RT, then 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.

To further reduce the number of sequences the set was checked for duplicates. 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 it was already present it was discarded from further consideration. If not, it was added to that list and checked for duplicates among the remaining sequences in the original set. If duplicates were 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-$\overline{l_i}$)/*number of duplicates* ($\overline{l_i}$ - average local confidence), whichever was higher. After that the average local confidence was recalculated as

$$\overline{P} = \sqrt[N]{\prod_1^N l_i}$$

where N is the length of the sequence.

Then an algorithm aimed at identifying sequences that were most probably from the same peptide was implemented. For each pair of sequences their every possible overlapping combination was checked using 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 $\overline{P} = \sqrt[n]{P}$ (n is the number of pairs used in comparison) was used to find the most probable overlap for the pair of sequences.

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 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 as usual

$$\overline{P} = \sqrt[N]{\prod_1^N l_i}$$

where N is the length of the new sequence.

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

Results and discussion

Despite the multiple preparative wash runs of the system the reconstructed by de novo sequences (around 4500) were first filtered to remove the remaining contaminants from previous runs – usual non-artificial peptide sequences identified by the swiss-prot-human database (about 250). The 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.

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 cutting down the number of repeating possible sequences and automatization became obvious.

First duplicates were removed from the set of sequences but the confidence levels of the sequences were updated to represent the higher reliability of those with multiple repeats. About 1000 duplicates were removed.

Then the filtered set was subjected to 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.

ANALYSLSREQ - 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?

SPECGTRUM - spectrum?

NTOFMATERLAL – (amou)nt of material

In case of a usual proteomic study the de novo and MS/MS results are searched in various proteome bases, such as SwissProt [2], to identify the originating protein or even organism, using MS specialized "search engines", such as MASCOT, Comet, Xtandem! and others. For this case another base and search engine were used — 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 books[1] 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 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 is 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.

To verify and sequence the less confident fragments the quote was converted into a FASTA file by deleting spaces and punctuation, replacing the missing letters by corresponding modifications and fed to PEAKS and Mascot.

Each language has its own patterns so in order to increase the reliability of the search a set of other quotes was added to the FASTA taken from: the same book of J.J. Thomson [4];

[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.

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 figure 1 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 (figure 1) – the end of the sentence is 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 an ellipsis 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 or its meaning is not explained by the challenge designers we suppose that it is an article and these missing words are thus “an infinitesimal”, which are totally uncovered in our data analysis. Considering these absences, the observed coverage reaches 90%.

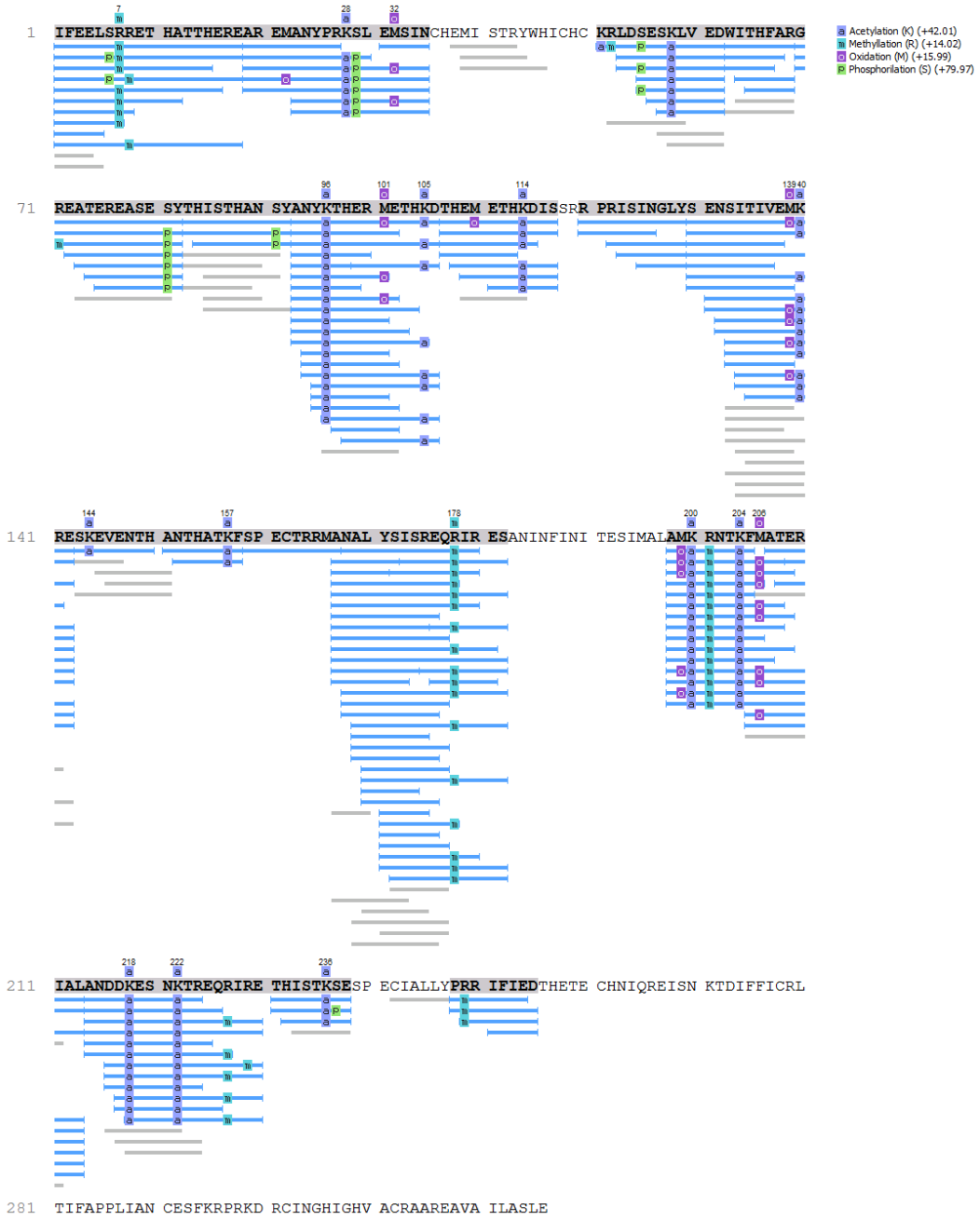
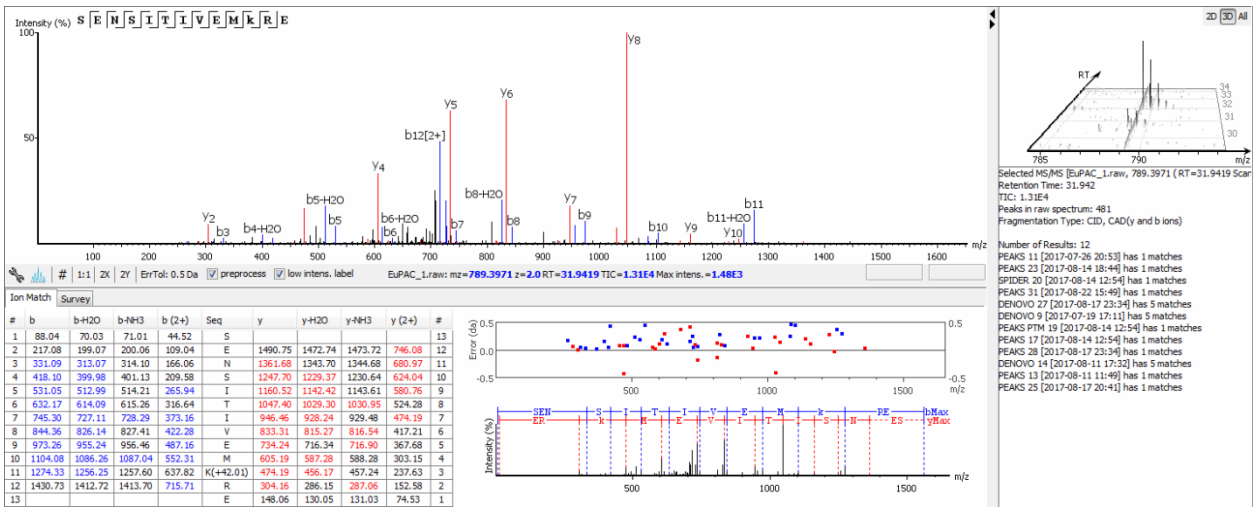


Figure 2. Sequence coverage of the original quote when only modifications from the sample description are allowed.



#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	88.04	70.03	71.01	44.52	S					13
2	217.08	199.07	200.06	109.04	E	1490.75	1472.74	1473.72	746.08	12
3	331.09	313.07	314.10	166.06	N	1361.68	1343.70	1344.68	680.97	11
4	418.10	399.98	401.13	209.58	S	1247.70	1229.37	1230.64	624.04	10
5	531.05	512.99	514.21	265.94	I	1160.52	1142.42	1143.61	580.76	9
6	632.17	614.09	615.26	316.64	T	1047.40	1029.30	1030.95	524.28	8
7	745.30	727.11	728.29	373.16	I	946.46	928.24	929.48	474.19	7
8	844.36	826.14	827.41	422.28	V	833.31	815.27	816.54	417.21	6
9	973.26	955.24	956.46	487.16	E	734.24	716.34	716.90	367.68	5
10	1104.08	1086.26	1087.04	552.31	M	605.19	587.28	588.28	303.15	4
11	1274.33	1256.25	1257.60	637.82	K(+42.01)	474.19	456.17	457.24	237.63	3
12	1430.73	1412.72	1413.70	715.71	R	304.16	286.15	287.06	152.58	2
13					E	148.06	130.05	131.03	74.53	1

Figure 3. Example of peptide sequencing results basing on MS/MS data.

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 also registered for other sequence fragments containing the C residue (“which”, “could”) and several contaminant peptides as well. Also the word “chemistry” de novo sequenced as starting with EE or sodiated EE, with a mass difference of 39.99, 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 (figure 4).

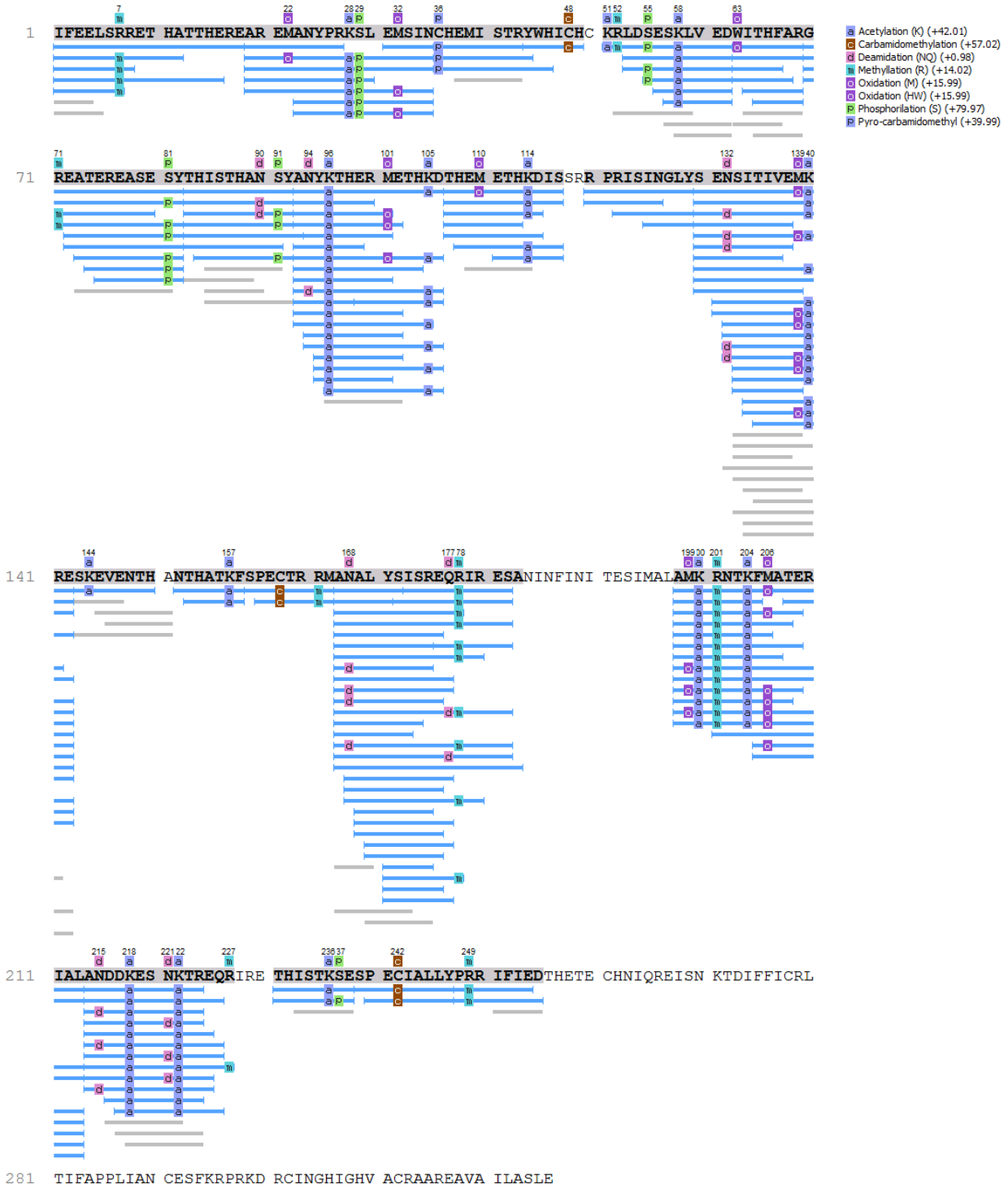


Figure 4. Sequence coverage change when additional identified modifications are allowed.

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 (figures 3 and 4). For reliability MALDI spectra were reacquired on a BRUKER Ultraflex instrument which allows better resolution and mass accuracy than its smaller brother. The peptides “could be solved”, “spectrum” and

“specially” were observed by all instruments only in their carbamidomethylated 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)TEDYLSLLLNR and subjected to a BLAST search which returned 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. Since these peptides were present on all 3 instruments, both MALDI targets in various spots 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 determine 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 carbamidomethylated.

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 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 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 lack 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 the 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 does not show coverage the corresponding peaks without such modification seem to be present in the sample spectra. So the question on the presence and localization of this modification remains open.

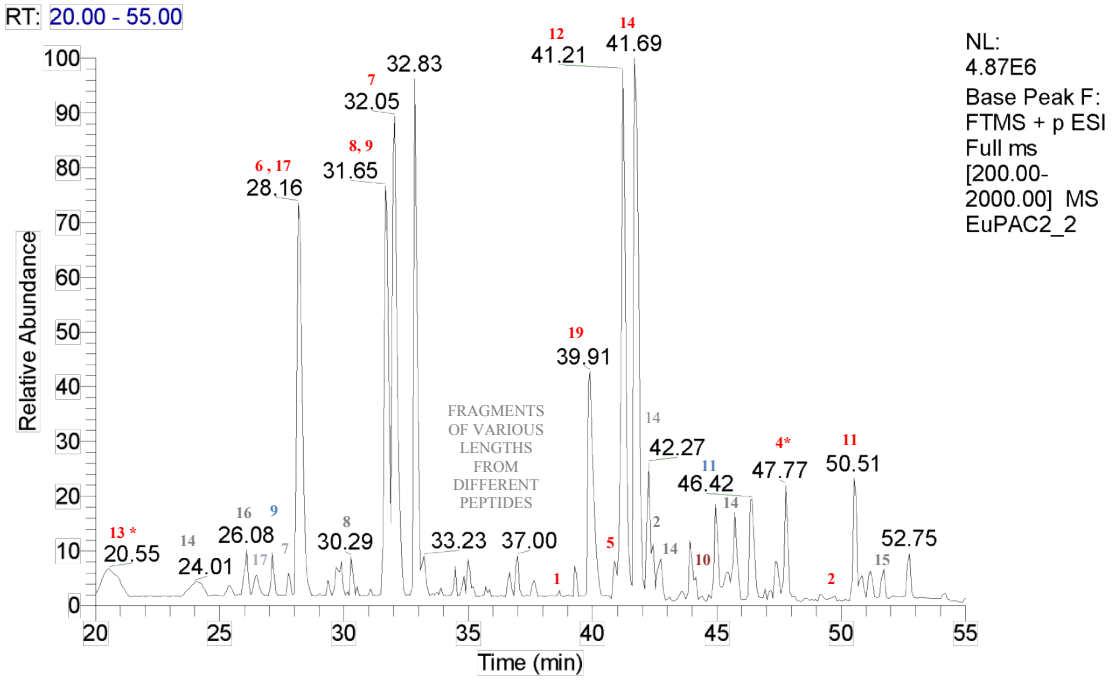


Figure 5. Annotation of LC peaks to peptides – * - carbamidomethylated, red – full peptide, grey – part of peptide, blue - oxidized peptide.

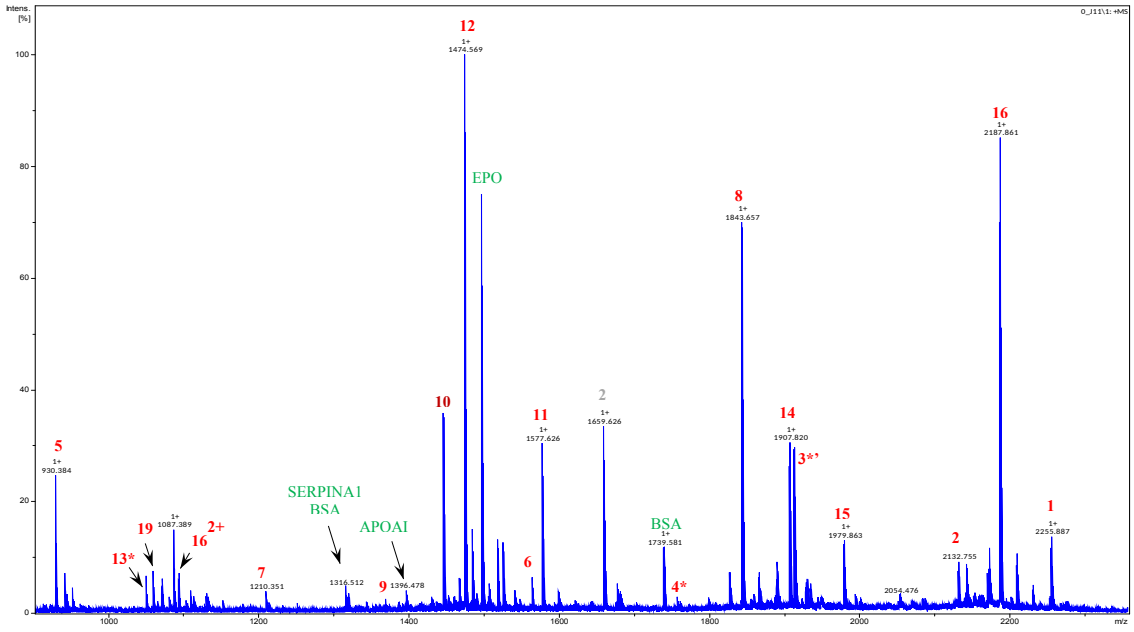


Figure 6. Annotation of MALDI peaks to peptides – * - carbamidomethylated, red – full peptide, grey – part of peptide, green – contaminants.

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.

Acknowledgements

Authors wish to thank EuPA for providing such an interesting exercise, Katerina Poverennaya (IBMC RAS) for her organization efforts for russians' 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.

Conflict of interest

The authors declare no competing financial interest.

List of publications

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 // Philosophical magazine, series 6 (1920), V.39, I.233, P.611-625
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. (2012), V.11
8. *M. Geisow*, Protemics: one small step for a digital computer, one giant leap for humankind, Nature Biotechnology (1998), V.16, I.2, P.206
9. *J.D. Watson and F.H.C. Crick* A Structure for Deoxyribose Nucleic Acid // Nature (1953) V.171, P.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"

The authors declare no competing financial interest.