EuPA Open Proteomics 22–23 (2019) 19–21

Contents lists available at ScienceDirect

EuPA Open Proteomics

journal homepage: www.elsevier.com/locate/euprot

Let me infuse this for you – A way to solve the first YPIC challenge

Britta Eggers, Sandra Pacharra, Martin Eisenacher, Katrin Marcus, Julian Uszkoreit (Dr.)*

Ruhr University Bochum, Faculty of Medicine, Medizinisches Proteom-Center, Gesundheitscampus 4, D-44801, Bochum, Germany

ABSTRACT

In a common proteomics analysis today, the origins of our sample in the vial are known and therefore a database dependent approach to identify the containing peptides can be used. The first YPIC challenge though provided us with 19 synthetic peptides, which together formed an English sentence. For the identification of these peptides, a *de-novo* approach was used, which brought us together with an internet search engine to the hidden sentence. But only having the sentence was not sufficient for us, we also wanted to identify as many as possible of the spectra in our data. Therefore, we created and refined a database approach from the de-novo method and finally could identify the peptide-sentence with a good overlap.

1. Introduction

The EuPA (European Proteomics Association) Young Proteomics Investigators Club (YPIC) prepared a challenge for its members. The task sounded very simple in the beginning: you will be provided by a solution of 19 synthetic peptides, which together form an English sentence. The participants of the challenge were free to choose the mass spectrometrical proteomics approach of their choice to find out this sentence and identify the peptides in the vial.

But the devil was in the detail: while most commonly a database approach in proteomics is used to identify the peptides, this was not possible to do here, as we had no known biological species representing "English language" and no hint, what sentence the peptides might build. Therefore, a less widely used *de-novo* approach had to be used for the spectra annotation.

Finally, with mixing *de-novo* and traditional database approaches, we were able to find the hidden sentence from a well-known book, though unfortunately we were not able to identify all peptides with our measurements.

2. Material and methods

2.1. Sample preparation

The provided synthetic peptides were kept in 30% acetonitrile (CAN). The description provided us with the information, that roughly 0.5 nmol/peptide were assigned in the mixture and in total 19 peptides were combined in the peptide mixture. Prior to MS analysis the peptide mixture was further diluted to ensure proper measurements and to prevent overloading of the column. Peptides

measurements were performed with 15 fold and 30 fold dilution of the sample, leading to a concentration of approximately 15.83 pmol/μL (15 fold) and 7,9 pmol/μL (30 fold).

2.2. Label free data dependent acquisition

The Nano HPLC analysis was performed on an UltiMate 3000 RSLC nano LC system (Dionex, Idstein, Germany) as described in [1]. The HPLC system was online-coupled to the nano ESI source of a Q Exactive HF mass spectrometer (Thermo Fisher Scientific, Germany). Full MS spectra were scanned in a range between 350 and 1,400 m/z with a resolution of 60,000 at 200 m/z for the detection of precursor ions (AGC target 3 × 106, 80 ms maximum injection time). The spray voltage was set to 1,500 V (+), and the capillary temperature to 275 °C. Lock mass polydimethylcyclosiloxane (445.120 m/z) was used for recalibration. The m/z values initiating MS/MS were set on a dynamic exclusion list for 30 s, and the top ten most intensive ions (charge state +2, +3, +4) were selected for fragmentation.

MS/MS fragments were generated by high-energy collision-induced dissociation (HCD) with a normalized collision energy (NCE) of 28%, fixed first mass of 100.0 m/z and an isolation window of

1.6 m/z. The fragments were analysed in an orbitrap analyser with a resolution of 30,000 at 200 m/z (AGC 1x106, maximum injection time 120 ms).

In total, we only used two LC–MS/MS measurements for the analysis of the YPIC samples.

2.3. Direct infusion analysis

Samples were loaded in a 250 μL Hamilton syringe and injected by a syringe pump (flow rate 3 μL/min) into the HESI source and were measured for 2.5 min with a full MS dd MS^2 method on a QExactive HF

* Corresponding author.
E-mail address: julian.uszkoreit@ruhr-uni-bochum.de (J. Uszkoreit).

https://doi.org/10.1016/j.euprot.2019.07.007
Received 9 July 2019; Accepted 17 July 2019
Available online 16 August 2019

mass spectrometer (Thermo Scientific)

In the ESI-MS/MS analysis, full MS spectra were scanned in a range between 350 and 1,400 m/z with a resolution of 60,000 at 200 m/z for the detection of precursor ions (AGC target 3 × 106, 80 ms maximum injection time). The spray voltage was set to 1,500 V (+), and the capillary temperature to 275 °C. Lock mass polydimethylcyclosiloxane (445.120 m/z) was again used for internal recalibration. The m/z values initiating MS/MS were set on a dynamic exclusion list for 5 s, and the top ten most intensive ions (all charge states except unassigned and 1) were selected for fragmentation experiments with an NCE of either 25%, 28% or 30% were used as well as a stepped gradient going from 26% to 27% up to 29% NCE.

For the YPIC analysis we used five of these direct infusion measurements.

2.4. Data analysis

By the given task and the fact, that the synthetic peptides were actual English words and no common peptides, it was clear we had to use a *de novo* spectrum identification method instead of the usual database searches. It might have been interesting to create a FASTA file for a complete English dictionary using the given translation code and special modifications for some letters, though finally we decided to use the free and open source tool DeNovoGUI [2] (version 1.15.11) for the data interpretation.

Prior to the actual analysis, the recorded RAW files were converted into mzML and MGF using ProteoWizard’s msConvert [3]. The resulting MGF of one of the intuitively best-looking LC–MS/MS run was fed into DeNovoGUI using the pNovo [4] and Novor [5] algorithms. As search parameters we initially set the strict default settings for the QExactive HF: parent mass tolerance of 5 ppm, fragment mass tolerance of 20 mmu. As we were told, that there might be some modifications which should also be interpreted as special letters in the final solution, we allowed the variable modifications acetylated lysine, phosphorylated serine and methylated arginine, besides the obligatory oxidation of methionine. With these settings, DeNovoGUI was able to identify 6616 spectra, of which the most identifications were rather bad.

Nevertheless, we went further and inspected the best hits visually. Here, it was good to know that we actually had only 19 peptides in the mixture, therefore in an ideal world we would have had only 19 different parent masses to inspect. Even though there were more masses in the results due to fragments and maybe synthesising artefacts, the number of these spectra was limited and therefore feasible for a first inspection. One thing we found out relatively fast was also the fact, that the Novor results seemed to be more accurate than the pNovo results. So, after sorting the results by the Novor score, we ended up in finding our first peptide: The spectra corresponding to 465.75 m/z had a good identification of the sequence WLTHFAR. As leucine and isoleucine are known to be indistinguishable by LC–MS/MS, some thinking brought us to the sequence WITHFAR. At this point, we set a threshold to identify at least five sequences by further inspection. For this, the results were sorted by m/z and Novor score and further inspected. After some time, we found four more, relatively well annotated sequences: SENSITIVEMkRE, SkEVENTHATkF, THEMETHkDIS and ANYkTHERMETHkD. These five sequences obviously translated to the English phrases “with far”, “sensitive more”, “so even that of”, “the method is” and “any other method”.

As we were stuck and our eyes hurt from inspecting spectra, we came up with the idea, that five of 19 peptides might be enough to find the actual sentence, if it is from a known source. Therefore, we used a well-known internet search engine, typed in the phrases and found the following sentences:

"I feel sure that there are many problems in chemistry, which could be solved with far greater ease by this than any other method. The method is surprisingly sensitive - more so than even that of spectrum analysis, requires an infinitesimal amount of material, and does not require this to be specially purified." These paragraph was from the book “Rays of Positive Electricity and Their Application to Chemical Analyses” by Sir J. J. Thomson [6].

The rest of the analysis was straightforward: Taking the sentences, all possible peptides with a length of 5–50 amino acids were created and put into a FASTA file for searching by X!Tandem [7] with the same settings as described for DeNovoGUI. As cleavage enzyme we set Trypsin, but with up to 10 allowed missed cleavages. Constraining the peptides and leaving out the parts, which we already had identified, we had 377 peptides in our FASTA database. We now searched all our MS/MS files, the LC based and the direct infusions, with this database and cut all identifications at the 0.01 X!Tandem expectation score. The peptides were further filtered to have at least 10 spectra per sequence. The longest continuous sequences were taken and added to the database, meaning if “ANALYSISREQ” and “ANALYSISREQRIRES” was found, only “ANALYSISREQRIRES” was added. With this, the original sentences were *in-silico* digested again, taking the newly found sequences as ground truth. This process was iterated five times, always enhancing the peptides in the database.

3. Results

With the described method of using X!Tandem and enhancing the database with peptide sequences, we finally ended up with a FASTA containing 19 entries, one for each expected peptide. Even though we found m/z traces in our data for all of these peptides, we could confidently identify only twelve of the peptides, three could be identified only sparsely with opening up the tolerances and one more only by fragments of the peptide (see Table 1). Three peptides could not be identified at all with our recorded data. Maybe, here other techniques like PRM or the injection of higher amounts could have helped.

With these identifications we could not completely cover the whole sentence or identify all 19 peptides in the provided solution. But finally we are rather sure, that the peptides were created to form the preamble of J. J. Thomson’s book.

4. Discussion

While we could identify the sentence and most of the peptides after we finally had the hint and a database, the identification using current

Table 1
The peptides which could finally be identified by MS/MS spectra. In bold the peptides, which were also spotted in the *de-novo* analysis are highlighted. The peptide on rank 13–15 (in italics) could only be identified after widening the parent and fragment mass tolerances.

Rank (by number of identified spectra)	Peptide	Text in original sentence
1	ANALYSISREQRIRES	ANALYSIS, REQUIRES
2	**SENSITIVEMKRE**	SENSITIVE - MORE
3	**SKEVENTHATKF**	SO than EVEN THAT OF
4	**THEMETHKDIS**	THE METHOD IS
5	**WITHFAR**	WITH FAR
6	ANDDKESNKTREQRIRE	AND DOES NOT REQUIRE
7	THISTKSE	THIS TO BE
8	PRRIFIED	PURIFIED
9	SYTHISTHAN	BY THIS THAN
10	**ANYKTHERMETHKD**	ANY OTHER METHOD
11	AMKRNTKFMATERIAL	AMOUNT OF MATERIAL
12	AREMANYPRKSLEMSIN	ARE MANY PROBLEMS IN
13	*SRRPRISINGLY*	SURPRISINGLY
14	*IFEELSRRETHATTHERE*	I FEEL SURE THAT THERE
15	*SPECIALLY*	SPECIALLY

de-novo software was rather disappointing. Even a retrospective analysis of the data was rather inconclusive and did not show us all peptides. Though one thing was striking to the eye: the best *de-novo* identified peptides were "tryptic" peptides, meaning the ones ending with an R or K, most prominently the sequence WITHFAR. This was most probably due to the fact, that the algorithm tries to take a tryptic digestion in the background and needs the resulting b-ion as a starting point. This then further hints, that the algorithms are actually not performing bad in real life data, but only had a hard time with the provided synthetic peptides. Another difficult task, the protein inference [8,9] from de-novo identified peptides, could not be applied in this challenge, but would also be something worth analysing in more depth.

5. Conclusions

Overall, the task to identify 19 purified non-tryptic synthetic peptides was not as easy as it seemed to be. We needed some visual inspection of the data and some refinement of databases to confidently identify only three of the peptides, and slightly identify another three of them. Nevertheless, the task given by the YPIC was a very interesting one and no one of the authors had

to try *de-novo* approaches before. Overall, we were happy to be able to find the hidden sentence in the peptides by only applying open source software and are looking forward to the next challenge.

Acknowledgements

The authors like to thank YPIC for the interesting challenge. Furthermore, we like to thank Christina Bunse for all the help with the data acquisition, especially regarding the direct infusion.

References

[1] J. Chen, S. Shinde, M.H. Koch, M. Eisenacher, S. Galozzi, T. Lerari, K. Barkovits, P. Subedi, R. Krüger, K. Kuhlmann, B. Sellergren, S. Helling, K. Marcus, Low-bias phosphopeptide enrichment from scarce samples using plastic antibodies, Sci. Rep. 5 (2015) 11438.

[2] T. Muth, L. Weilnböck, E. Rapp, C.G. Huber, L. Martens, M. Vaudel, H. Barsnes, DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra, J. Proteome Res. 13 (2) (2014) 1143–1146.

[3] M.C. Chambers, B. Maclean, R. Burke, D. Amodei, D.L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T.A. Baker, M.Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S.L. Seymour, L.M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E.W. Deutsch, R.L. Moritz, J.E. Katz, D.B. Agus, M. MacCoss, D.L. Tabb, P. Mallick, A cross-platform toolkit for mass spectrometry and proteomics, Nat. Biotechnol. 30 (10) (2012) 918–920.

[4] H. Chi, R.X. Sun, B. Yang, C.Q. Song, L.H. Wang, C. Liu, Y. Fu, Z.F. Yuan, H.P. Wang, S.M. He, M.Q. Dong, pNovo: de novo peptide sequencing and identification using HCD spectra, J. Proteome Res. 9 (5) (2010) 2713–2724.

[5] B. Ma, Novor: real-time peptide de novo sequencing software, J. Am. Soc. Mass Spectrom. 26 (11) (2015) 1885–1894.

[6] J.J. Thomson, Rays of Positive Electricity and Their Application to Chemical Analyses, (1913).

[7] R. Craig, R.C. Beavis, TANDEM: matching proteins with tandem mass spectra, Bioinformatics 9 (2004) 1466–1467.

[8] J. Uszkoreit, A. Maerkens, Y. Perez-Riverol, H.E. Meyer, K. Marcus, C. Stephan, O. Kohlbacher, M. Eisenacher, PIA: an intuitive protein inference engine with a web-based user interface, J. Proteome Res. 14 (7) (2015) 2988–2997.

[9] J. Uszkoreit, Y. Perez-Riverol, B. Eggers, K. Marcus, M. Eisenacher, Protein inference using PIA workflows and PSI standard file formats, J. Proteome Res. 18 (2) (2019) 741–747.