Contents lists available at ScienceDirect

EuPA Open Proteomics

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

Quick and clean: Cracking sentences encoded in *E. coli* by LC–MS/MS, de novo sequencing, and dictionary search

Lili Niu[a], Matthias Mann[a,b,*]

[a] Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
[b] Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

ARTICLE INFO

Keywords:
EuPA YPIC challenge
Mass-spectrometry
Proteomics
De novo sequencing
Dictionary search

ABSTRACT

In this study, we faced the challenge of deciphering a protein that has been designed and expressed by *E. coli* in such a way that the amino acid sequence encodes two concatenated English sentences. The letters 'O' and 'U' in the sentence are both replaced by 'K' in the protein. The sequence cannot be found online and carried to-be-discovered modifications. With limited information in hand, to solve the challenge, we developed a workflow consisting of bottom-up proteomics, de novo sequencing and a bioinformatics pipeline for data processing and searching for frequently appearing words. We assembled a complete first question: "Have you ever wondered what the most fundamental limitations in life are?" and validated the result by sequence database search against a customized FASTA file. We also searched the spectra against an *E. coli* proteome database and found close to 600 endogenous, co-purified *E. coli* proteins and contaminants introduced during sample handling, which made the inference of the sentence very challenging. We conclude that *E. coli* can express English sentences, and that de novo sequencing combined with clever sequence database search strategies is a promising tool for the identification of uncharacterized proteins.

1. Introduction

Today, protein identifications in mass-spectrometry (MS)-based shotgun proteomics mostly rely on sequence database search or spectral library searching, which require a protein sequence database as prior knowledge [1]. Thus, these approaches are limited when analyzing tandem mass spectra derived from uncharacterized proteins or unknown species. Conversely, de novo sequencing algorithms are well suited to this purpose, and are increasingly empowered by the ever-improving resolving power and mass accuracy of mass spectrometers [2]. Nevertheless, it is very difficult to be sure about the exact sequence of a peptide by de novo sequencing alone. This typically requires peptides of a certain make up, very good MS/MS data quality and perhaps different fragmentation methods. To obtain nearly correct overall sequences, or very accurate sub-sequences that can be assembled into peptide sequence tags [3], is much more feasible. It is such an 'error tolerant' approach that we use here to tackle the challenge.

2. Results and discussion

2.1. Formation of the strategy

In this year's YPIC Challenge, the organizers offered several Challenge categories, including answering *E. coli's* question, three-dimensional Grammar, bioinformazing, protein punctuation, and bioreactivity (http://eupa.org/ypic/the-challenge/). We were particularly interested in developing a workflow to decode the protein sequence encoding a question in *E. coli*. We received the dry protein three weeks before the manuscript submission deadline. After thinking about all the proteomics workflows that we could apply to this challenge, we decided to go for a bottom-up proteomics approach followed by de novo sequencing. In each of these two modules, there are aspects that can be taken into consideration to improve them. The use of multiple proteases is obviously attractive, since they cleave peptides at different sites, creating a diversity of peptides, hence in principle this should increase the proteome sequence coverage and allow an easier assembly of neighboring peptides. However, pressed by time and due to limited availability of proteases, we decided to digest the protein by trypsin and LysC, the most commonly used enzymes in proteomics sample preparation. We purified the peptides on Stage Tips with washing buffers

* Corresponding author at: Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark.
E-mail addresses: lili.niu@cpr.ku.dk (L. Niu), mmann@biochem.mpg.de (M. Mann).

https://doi.org/10.1016/j.euprot.2019.07.010
Received 9 July 2019; Accepted 17 July 2019
Available online 29 July 2019

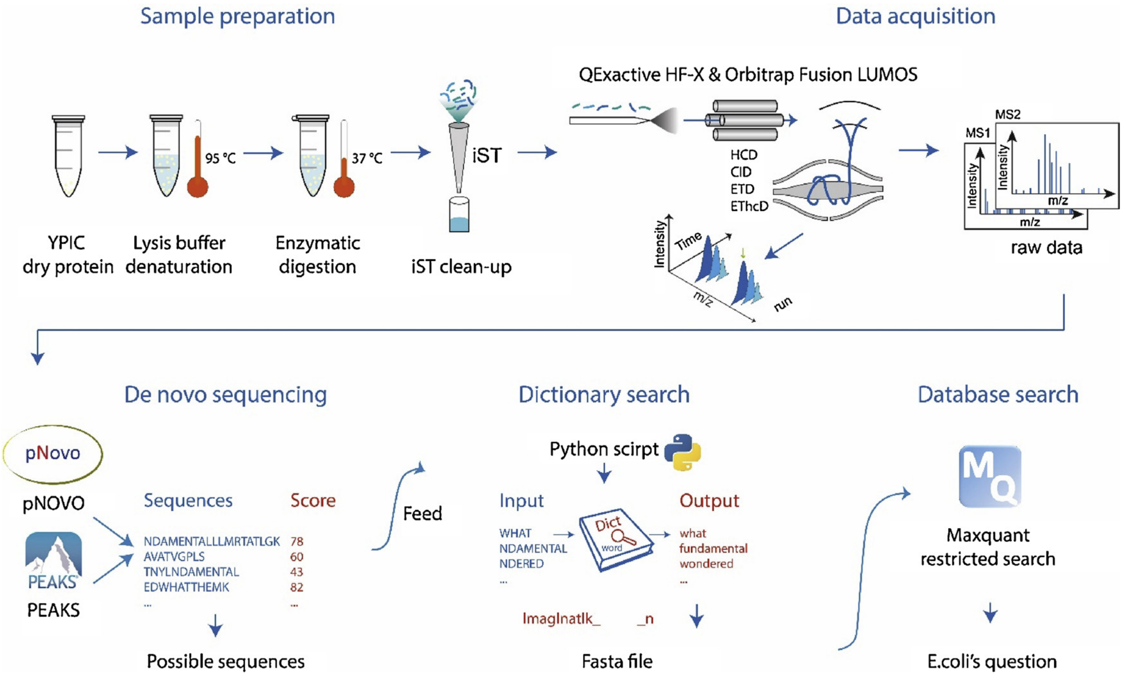

Fig. 1. Analysis workflow.

from PreOmics (PreOmics GmbH, Martinsried, Germany) and this resulted in around 10 µg (80%) of peptides as determined by Nanodrop, a promising amount for our efforts at deciphering the sentence.

2.2. Initial excitement – de novo sequencing

Having the peptides ready, we analyzed the peptide mixture with the state of the art LC–MS/MS workflow of our laboratory in order to get a first impression of the sample. After the acquisition was complete, we visually inspected the raw file. The total ion chromatogram looks quite reasonable in terms of equal distributions of peptide signals with a handful of major peaks throughout the LC gradient. We then analyzed the raw file in the pNovo software (version 3), a free-access de novo peptide sequencing tool [4]. This analysis resulted in 74, 431 possible peptide sequences with length ranging from two to 34 amino acids and peptide spectrum match (PSM) scores of 0 to 114. Sorting the sequences based on PSM score, two words – 'mental' and 'fear' – already appeared in the highest scoring sequence [NDAMENTALLLMRTATLGKNSLNLLFEAR]. With this initial excitement, we went on and started looking for other words in the remaining sequences one by one. Our excitement and patience quickly ran out, however, as not a single further word appeared, and the prospect of analyzing 70,000 sequences by hand was daunting. Clearly, we needed an automated tool to extract words from the sequence list. At this point, we concluded that the next step would be to develop a 'dictionary search tool'. At the same time, we decided to analyze the sample on a different instrument – Orbitrap Fusion Lumos (Thermo Fisher) – to take advantage of multiple fragmentation modes (CID, HCD, ETD and EThcD) since we had plenty of sample left.

2.3. Developing a dictionary search tool

Similar to the concept of spectrum library search, we decided to write a Python script that searches a sequence list against a library of words and reports the matched words along with the frequency of their occurrences. This frequency to some extent indicates the likelihood of the word (or the sequence from which the word is derived) being indeed present in the sample.

Considering that the dictionary should not be too large but still sufficient to cover the commonly used words, we found a list of the top 10, 000 most frequent English words from Github (https://github.com/first20hours/google-10000-english). Also considering that the protein encodes two questions, "both touching on the same fundamental biological issue", we decided to generate another dictionary that covers biology-related terms. The first approach coming to mind was to extract words from a biological textbook. We chose Molecular Biology (5th Edition) by Robert F. Weaver from which we extracted a total of 32, 222 word entries. Out of curiosity, we checked the most frequently appearing words in this book apart from prepositions, conjunctions and pronouns. "DNA", "RNA", "gene", "protein" and "transcription" are the top five ones, perhaps not surprisingly. Interestingly, the word "colleagues" ranks 66, appearing 948 times, right after the word "cells", indicating the importance of teamwork in science. These two dictionaries overlapped by 4344 words, and together they have 37,748 entries. We also took special care of the words that contain "O", "U" and "R", such that they were further sliced into two 'half-words' according to the cleavage pattern of trypsin, while also keeping intact words for the situation of mis-cleavage. With this, the combined dictionary grew to 57, 016 entries. At this point, we had a dictionary of words, and we had generated a script that does the search (Fig. 1).

2.4. Extracting words and assembling sentences

We also took special care on the peptide sequence side. Considering the instructions that in the sentence "O and U are both replaced by K in the protein", we duplicated all sequences that contain "K" by substituting "K" for either "O" or "U". We also duplicated the sequences that contain "L" by substituting "L" for "I" while keeping the rest unchanged. Eager to test the script out, we ran the analysis and this resulted in a *.csv* file which contained all found words along with their frequencies. The first clue we got from this was the word "fundamental". We then manually search for the sequences and inspect the corresponding MS2 spectra in both pNovo and PEAKS Studio (see Material and Methods section) to make sure that this is the complete word, based on PSM score, precursor peak shape and matched b-, and y-ions. With this approach, we consecutively found the words "have", "ever", "wondered", "what", "_tation", and more. After filtering by inspecting spectra and looking for neighboring sequences based on words we are familiar with, we found evidence for the sentence "Have you

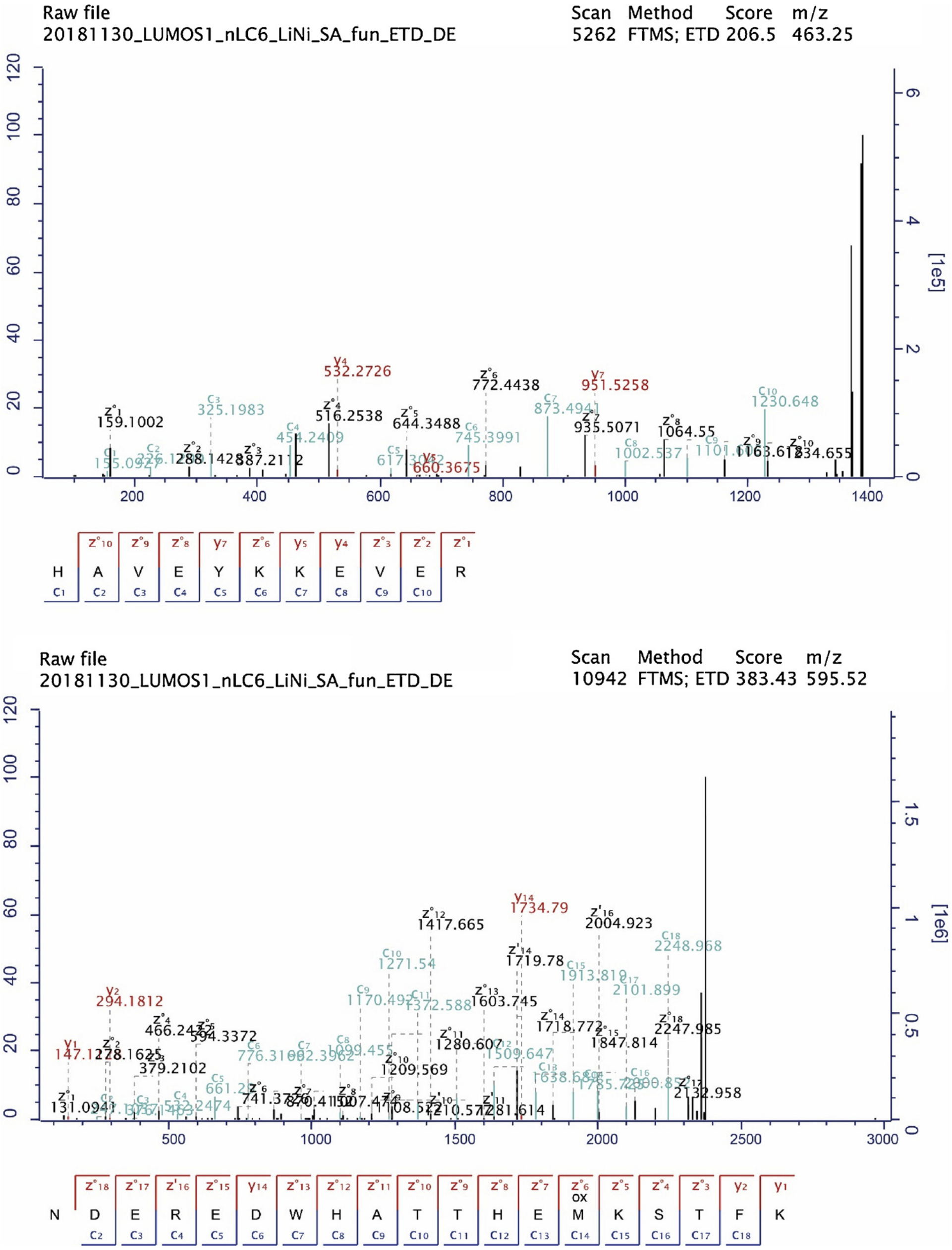

Fig. 2. MS/MS spectra of sequences generated on Orbitrap Fusion Lumos under ETD mode.

ever wondered what the most fundamental limitations in life are". We also found that this should be the first sentence as there were matched spectra for the sequence "AGRHAVEYK" which contains the beginning tag "AGR". Unfortunately, we only found pieces of the second sentence - "Is there a…mes to what you…" (Fig. 4a).

2.5. Validation by restricted search

To validate our finding, we generated a FASTA file containing the sentence as well as the reference proteome of *E. coli* (strain K12, version 201812) and performed a standard sequence database search in MaxQuant software [5,6]. At this point, we had also collected raw files generated from Oribitrap Lumos Fusion. We took advantage of the versatile fragmentation modes available on that instrument apart from the higher-energy C-trap dissociation (HCD) that we use in our daily experiments on the Q Exactive HF-X instrument. The combination of different fragmentation modes should increase peptide identification, as certain peptides would favor one fragmentation mode over another. Electron transfer dissociation (ETD) for instance, induces amide bond break along the peptide backbone, producing complementary c- and z-type fragment ions, instead of b- and y-type ions in the case of HCD and CID. We manually inspected the MS2 spectra of all identified peptides of the sentence, and could validate all peptides with good spectra quality, shown by four example sequences in (Figs. 2 and 3). Interestingly, the best spectra of these sequences come from different fragmentation modes, for instance the sequences [HAVEYKKEVER] and [NDEREDWHATTHEMKSTFK] had the best score from the LUMOS instrument using ETD (Fig. 2). The sequences [NDAMENTALLLMLTATLK] had the highest score generated under CID mode on the same instrument, whereas the sequence [NSLNLLFEAR] was best from Q Exactive HF-X under HCD mode (Fig. 3).

In total, we found 23 peptides belonging to the sentence. The intensities of these 23 peptides span four orders of magnitude. Apart from these peptides, we also identified an additional 3606 peptides that are derived from either contaminants such as keratins during sample handling or endogenous proteins of *E. coli*, indicating that the sample

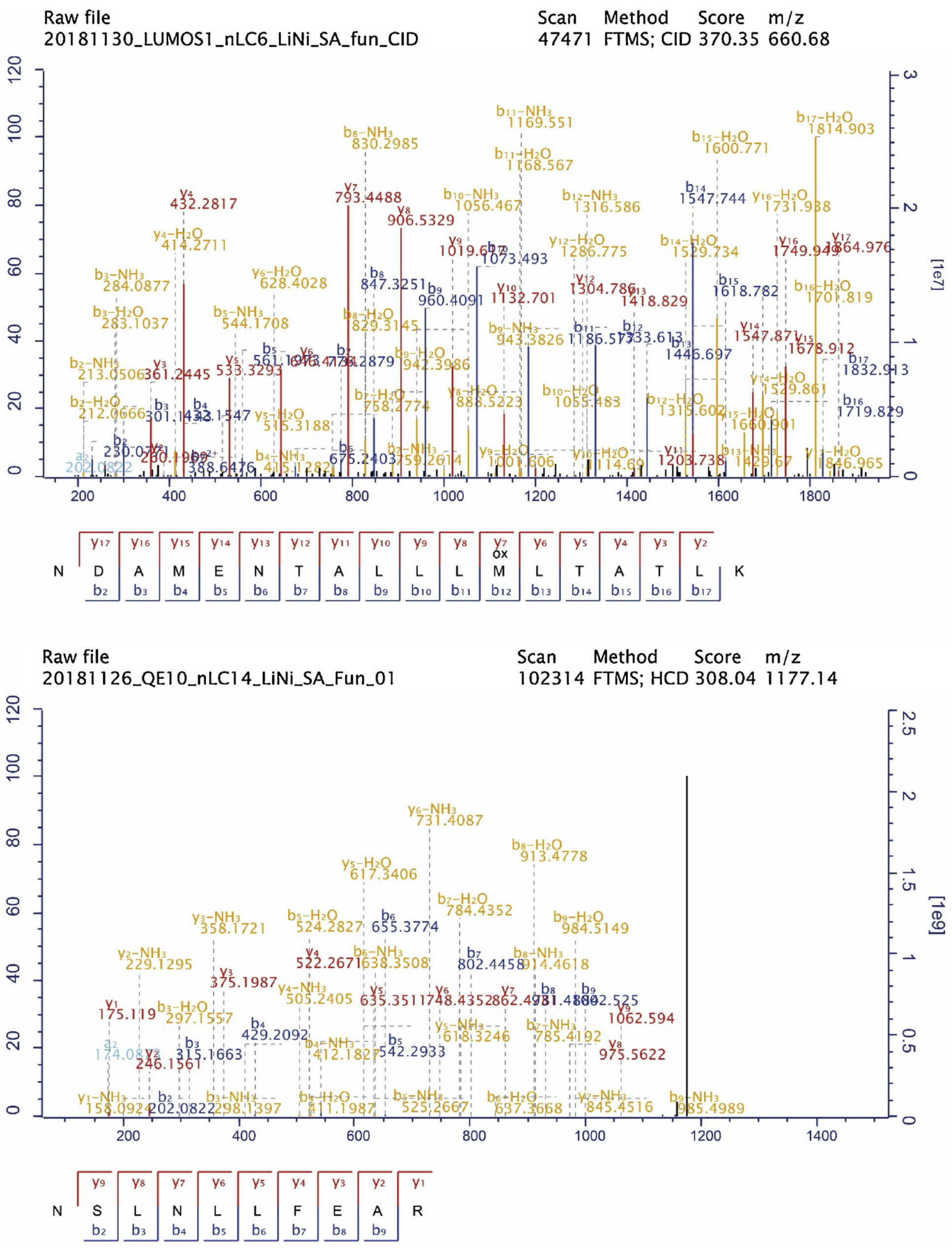

Fig. 3. MS/MS spectra of sequences generated on Orbitrap Fusion Lumos under CID mode and on Q Exactive HF-X under HCD mode.

was quite complex, when looking at it with modern proteomics methods. A good proportion of these peptides was even more abundant than our target peptides and only 10% among the top 100 ranking peptides belong to the sentence (Fig. 4b). All 3629 peptides were further assembled into around 591 protein groups, among which the one with the sentence has the highest intensity (Fig. 4c).

2.6. Answering *E. coli's* question

Our workflow has limitations, just as life itself. "Have you ever wondered what the most fundamental limitations in life are?" Dear *E. coli*, our answer would be "Yes, we have". Life in a way is limited by its form, trapped in a physical body by which the will executes actions. A physical body renews itself but there are limitations to this. It wears out, in the form of aging and diseases. Life is dependent and in most scenarios, it needs help from other organisms, and that is symbiosis. If the dependence is disrupted, life loses balance. As an example, the disruption of gut microbiota can cause us many troubles. Nonetheless, life is limited and because of that, life is precious.

3. Conclusion

In this study, we took advantage of mass spectrometric and bioinformatics tools established or under development in the field of proteomics and partially addressed the challenge of decoding *E. coli's* question within two weeks (including writing this manuscript). We conclude that bottom-up proteomics, de novo sequencing and constrained sequence database search is a promising pipeline for the identification of uncharacterized protein. Alternative fragmentation modes to HCD yield complementary information on fragment ions and generate better spectra quality in certain sequences.

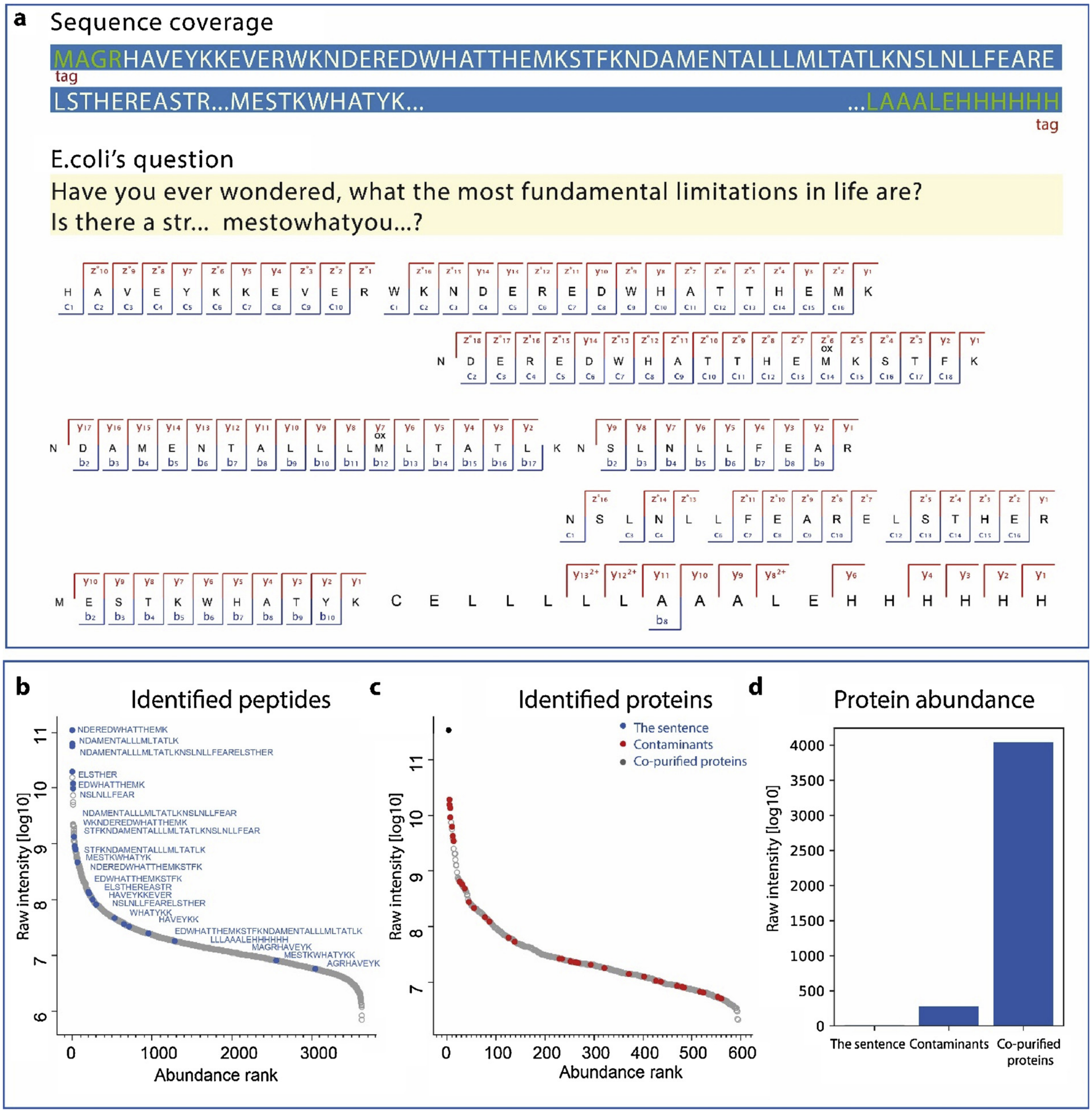

Fig. 4. Validation of found sentences by sequence database search.
a. Assembled peptide sequence, corresponding sentence and matched fragment ions of each sequence. b. All identified peptides from the sample. Peptides belonging to the sentence were highlighted in blue. c. All identified proteins from the sample, with the sentence, contaminants and co-purified proteins of *E. coli* color-coded. d. Summed raw intensity of the sentence, contaminants and co-purified proteins from *E. coli*.

4. Materials and methods

4.1. Sample preparation

Dry protein was dissolved in 40 µl of SDC reduction and alkylation buffer (PreOmics GmbH, Martinsried, Germany) and boiled for 10 min at 95 °C while vortexing at 1200 rpm to denature the protein [7]. The lysate was digested for 5 h with LysC and trypsin (0.25 µg each) at 37 °C and 1200 rpm. Peptides were acidified to a final concentration of 0.1% trifluoroacetic acid (TFA) to quench the digestion reaction. Peptide concentration was estimated using Nanodrop and sample was loaded on two 14-gauge Stage-Tip plugs. Peptides were washed first with isopropanol/1% TFA (200 µl) and then 0.2% TFA (200 µl) using a centrifuge at 2000xg. Peptides were eluted with 60 µl of elution buffer (80% acetonitrile/1% ammonia) and dried at 60 °C using a SpeedVac centrifuge (Eppendorf, Concentrator plus). Dried peptides were redissolved and sonicated in 20 µl of 5% acetonitrile/0.1% TFA and concentration measured using the Nanodrop.

4.2. LC–MS/MS

The sample was measured using LC–MS instrumentation consisting of an EASY-nLC 1200 system interfaced on-line with a Q Exactive HF-X Orbitrap or Orbitrap Fusion Lumos (all from Thermo Fisher Scientific). Purified peptides were separated on 50 cm HPLC-columns (ID: 75 µm; in-house packed into the tip with ReproSil-Pur C18-AQ 1.9 µm resin (Dr. Maisch GmbH)). Around 500 ng peptides were injected. Peptides were loaded in buffer A (0.1% formic acid) and eluted with a linear 82 min gradient of 3–23% of buffer B (0.1% formic acid, 80% (v/v) acetonitrile), followed by a 8 min increase to 40% of buffer B. The gradients then increased to 98% of buffer B within 6 min, which was kept for 4 min. Flow rates were kept at 350 nl/min. Re-equilibration was done for 4 µl of 0.1% buffer A at a pressure of 980 bar. Column temperature was kept at 60 °C using an integrated column oven (PRSO-V2, Sonation, Biberach, Germany). Two acquisition methods were used on the Q Exactive HF-X Orbitrap. Both were comprised of a Top15 data-dependent MS/MS scan (DDA, topN method). Target value for the full scan MS spectra was 3E6 in the 300-1,650 *m/z* range with a maximum injection time (IT) of 25 ms and a resolution of 60,000 at *m/z* 200. Precursor ions targeted for fragmentation were isolated with an isolation width of 1.4 *m/z*, followed by higher-energy collisional dissociation (HCD) with a normalized collision energy of 27 eV. Precursor dynamic exclusion was activated for a duration of 30 s. MS/MS scans were performed at a resolution of 15,000 at *m/z* 200 with an automatic gain control (AGC) target value of 1E5 and an IT of 25 ms or 50 ms.

Four acquisition methods were used on the Orbitrap Fusion Lumos instrument, differing in fragmentation modes – CID, HCD, ETD and EThcD. All methods used the same LC gradient as in the Q Exactive HF-X. AGC target for the full scan MS spectra was 4E5 in the 350-1,000 *m/z* range with an IT of 50 ms and a resolution of 120,000 at *m/z* 200. Precursor ions targeted for fragmentation were isolated with an isolation window of 1.6 *m/z*, followed by either CID (collision energy (CE) 30%), HCD (CE 35%), ETD (with the option "use calibrated charge-dependent ETD parameters" activated) or EThcD (CE 35% for HCD). Precursor dynamic exclusion was activated for a duration of 20 s. MS/MS scans were performed at a resolution of 30, 000 at *m/z* 200 with an AGC target of 1E5 and an IT of 54 ms.

4.3. *De novo sequencing*

De novo sequencing was performed with a publically available software pNovo and the commercial software PEAKS Studio. The software RawConverter was used to convert RAW files to mgf format to be analyzed in the pNovo software. In the analysis by pNovo, the enzyme specificity was set to Trypsin KR C, mass error tolerance for both precursor and fragment was set to ± 10 ppm, and the open search option was activated. The top 10 results were kept for each spectrum. In the analysis of PEAKS software, default settings were used except that precursor mass error tolerance was set to ± 10 ppm, and the search included cysteine carbomidomethylation as fixed modification, oxidation on methionine and N-terminal acetylation as variable modifications.

4.4. *Restricted search*

Restricted search was performed by MaxQuant v.1.5.3.30 software [5] using the integrated Andromeda Search engine [6] and pFind software, searching against a customized fasta file which contained the assembled English sentence. Enzyme specificity was set to trypsin with a maximum of 2 missed cleavages and the search included cysteine carbamidomethylation as fixed modification and oxidation on methionine and N-terminal acetylation as variable modifications with a minimum length of 7 amino acids. A false discovery rate (FDR) of 1% was set to PSM and protein levels.

4.5. *Bioinformatics analysis*

Bioinformatic analysis was performed in Python within a Jupyter Notebook. Identified peptide sequences from de novo sequencing by pNovo software were filtered for a minimum sequence length of 4 and spectrum score of 30. Sequences identified by PEAKS software were filtered for a minimum peak area of 10E4, an absolute maximum mass error of 5 ppm and a minimum average local confidence score (ALC) of 60%. The resulting sequences were then searched against a dictionary that contained 57, 016 words.

4.6. *Data availability*

MS proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD012015. Jupyter Notebook has been deposited to GitHub (https://github.com/llniu/YPIC_Challenge_2019).

Funding

The work carried out in this project was supported by The Novo Nordisk Foundation for the Clinical Proteomics group (grant NNF15CC0001) and The Novo Nordisk Foundation for the Copenhagen Bioscience PhD Program (NNF16CC0020906).

Author contributions

LN designed and performed the study, wrote the manuscript. MM edited the manuscript.

Declaration of Competing Interest

None.

Acknowledgments

We thank all members of the Clinical Proteomics group, and the Proteomics and Signal Transduction Group (Max-Planck Institute), in particular Alberto Santos Delgado for his inspiration, Atul Shahaji Deshmukh, Niels Skotte and Rajat Gupta for the good company. We thank Jeppe Madsen, Martin Rykær, and Simone Schopper for technical assistance. We thank the organizers of this challenge, the European Proteomics Association (EuPA), the Young Proteomics Investigator Club (YPIC) for this interesting event.

References

[1] J. Cox, M. Mann, Quantitative, high-resolution proteomics for data-driven systems biology, Annu. Rev. Biochem. 80 (2011) 273–299.
[2] R. Aebersold, M. Mann, Mass-spectrometric exploration of proteome structure and function, Nature 537 (2016) 347–355.
[3] M. Mann, M. Wilm, Error-tolerant identification of peptides in sequence databases by peptide sequence tags, Anal. Chem. 66 (1994) 4390–4399.
[4] H. Yang, H. Chi, W.J. Zhou, W.F. Zeng, K. He, C. Liu, R.X. Sun, S.M. He, Open-pNovo: De Novo peptide sequencing with thousands of protein modifications, J. Proteome Res. 16 (2017) 645–654.
[5] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, Nat. Biotechnol. 26 (2008) 1367–1372.
[6] J. Cox, N. Neuhauser, A. Michalski, R.A. Scheltema, J.V. Olsen, M. Mann, Andromeda: a peptide search engine integrated into the MaxQuant environment, J. Proteome Res. 10 (2011) 1794–1805.
[7] N.A. Kulak, G. Pichler, I. Paron, N. Nagaraj, M. Mann, Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells, Nat. Methods 11 (2014) 319–324.