QuantumSi

APP NOTE

Comparison of Protein Sequencing Analysis of CDNF, IL6, and FGF2 on Platinum™ and Mass Spectrometry

Introduction

The rapid growth of the field of proteomics since the early 2000s has been closely tied to the technological advancements in mass spectrometry (MS). This technique is used to identify proteins and is commonly performed as a service at core facilities due to the high cost, space, and expertise required. A typical workflow for protein identification via MS involves digestion of proteins into peptides followed by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) on a high-resolution analyzer. In the first MS stage (the precursor ion scan also known as MS1), peptides are separated by reversed phase chromatography and the mass-to-charge ratio is detected as intact peptides elute over time. In the second MS stage (the product ion scan also known as MS2), peptides are fragmented by collision-induced dissociation, generating spectra that can be analyzed to determine peptide sequence.[1] Complex software is used to map the spectra to a large database of simulated spectra from the proteome generated via *in silico* protein digestion to identify peptides and determine the protein of origin.

While MS has become the gold standard for protein identification, several confounding factors can limit the unambiguous, proteome-wide mapping of peptides. For example, missed or unanticipated cleavages, chemical modifications, and post-translational modifications (PTMs) can lead to peptides that are not detected by the search algorithm. Amino acids of identical or similar masses also present challenges in standard MS workflow. Examples include the differentiation of isoleucine and leucine, asymmetric dimethylarginine and symmetric dimethylarginine, and trimethyllysine and acetyllysine. Furthermore,

Q-Si Technology

Quantum-Si's benchtop Platinum™ instrument enables protein sequencing from biological samples in a simple user-friendly workflow. Our technology utilizes dye-tagged N-terminal amino acid recognizers and semiconductor chip technology to detect the binding characteristics and binding order of N-terminal amino acids, resulting in unique kinetic signatures that can be used to differentiate and identify amino acid residues and PTMs. A more detailed overview of the workflow and technology can be found in our Science Paper.

chemical modifications may result from LC-MS ionization, such as the modification from glutamine to pyroglutamate. These changes could be missed when using default databases and analysis settings.

In addition, the workflow is laborious and requires expensive capital equipment and advanced expertise in analysis to properly identify the protein sequence. Researchers often resort to simpler proteomic techniques such as western blots to detect protein changes, with the tradeoff of not uncovering deeper insights at the amino acid and peptide level of proteins.

Quantum-Si's single-molecule protein sequencing workflow on Platinum™ overcomes these challenges by offering an accessible and convenient benchtop instrument with Cloud-based analysis software that enables researchers to discover deeper proteomic insights without the need for expensive equipment and expertise. Proteins are digested with a protease to generate peptide libraries and immobilized on a semiconductor chip in as little as 2–3 hours of hands-on time. Protein sequencing occurs on the Platinum™ in 10 hours or less without the need for complex cyclical chemistry and fluidics. Sequencing data is automatically and securely transferred to a Cloud-based software environment for analysis. Protein identification output is provided without the need for expert analyses, enabling researchers to quickly make decisions about their protein samples.

Protein sequencing on Platinum offers several advantages for protein identification relative to tandem MS, including the convenience of a benchtop sequencer without the need to send samples to a core facility, the simplicity of the workflow without the need for advanced expertise, and the ability to gain deeper insights into the proteome thanks to the information-rich binding signatures of N-terminal amino acid (NAA) recognizers.

To compare protein sequencing to tandem MS for the purpose of identification of an isolated protein, we sequenced the proteins CDNF, FGF2, and IL6 on Platinum and sent the same proteins to an MS core facility for analysis.

Methodology & Workflow

Proteins are sequenced on Platinum using the Library Preparation Kit, the Protein Sequencing Kit, and our real-time dynamic sequencing workflow as previously described.[2] Briefly, proteins are digested into peptide fragments and conjugated to macromolecular linkers. The conjugated peptides are then immobilized on Quantum-Si's semiconductor chip with exposed N-termini for sequencing. Dye-labeled recognizers bind on and off to NAAs, generating pulsing patterns with characteristic fluorescence and kinetic properties. Regions corresponding to NAA recognition are termed recognition segments (RSs). Aminopeptidases in solution sequentially remove individual NAAs to expose subsequent amino acids for recognition. Fluorescence lifetime, intensity, and kinetic data are collected in real time and analyzed to distinguish amino acids of the peptide sequence.

Sequencing profiles of peptides are visualized as kinetic signature plots—simplified trace-like representations of the time course of complete peptide sequencing containing the median pulse duration (PD) for each RS and the average duration of each RS and non-recognition segment (NRS). When recognizers bind to NAAs, they also make important contacts with nearby downstream residues in the RS that influence the average PD of binding events between recognizers and target peptides. This kinetic sensitivity to nearby downstream residues provides a wealth of information on peptide sequence composition and is extremely beneficial for mapping traces from peptides to their proteins of origin.

In this study, recombinant human proteins of the cerebral dopamine neurotrophic factor (CDNF, 161 amino acids, R&D Systems, Cat# 5097-CD-050), fibroblast growth factor 2 (FGF2, 146 amino acids, R&D Systems, Cat# 233-FB), and Interleukin-6 (IL6, 204 amino acids, Cayman Chemical, Cat# 30173) were used as model proteins to demonstrate protein identification and single amino acid changes from sequencing data based on our kinetic model and proteome mapping software. We sequenced CDNF, FGF2, and IL6 using a set of five NAA recognizers [PS610 recognizes N-terminal phenylalanine (F), tyrosine (Y), and tryptophan (W); PS1223 recognizes leucine (L), isoleucine (I), and valine (V); PS1220 recognizes arginine (R); PS1259 recognizes N-terminal glutamine (Q) and asparagine (N); and PS1165 recognizes N-terminal alanine (A) and serine (S)]. This set of five recognizers recognizes a total of 11 NAAs. We digested CDNF, FGF2, and IL6 using the endopeptidase Lys-C and prepared a peptide library for sequencing. Data was collected and analyzed using our cloud-based software.

The same full-length CDNF, FGF2, and IL6 proteins were sent to an MS core facility at an academic institution for Lys-C digestion and LC-MS/MS analysis on the Q Exactive™ HF-X Hybrid Quadrupole-Orbitrap™ Mass Spectrometer equipped with dual pump UltiMate 3000 RSLCnano (Thermo Fisher, San Jose, CA). The UniProt human protein database was used as a reference for peptide identification.

Results & Discussion

The results from both MS and Platinum protein sequencing platform correctly identified the CDNF protein, despite different workflows and identification of different peptides (Figure 1).

The workflow for submitting samples to the MS core was carried out by multiple people and required shipment of samples to the core facility, processing of the samples followed by analysis of the data by the core facility. Sample submission involved scheduling time with the core facility, preparing the samples, and shipping the samples to the core. Once received, the samples were processed, digested, and run on MS equipment

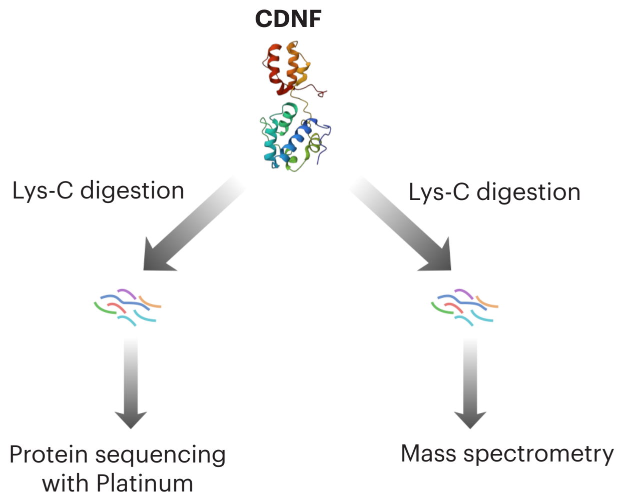

Figure 1. *CDNF Protein Sequencing with Platinum and Mass Spectrometry Workflow*

by highly trained staff. The data was then processed using a suite of software tools by the core facility and sent to Quantum-Si in the form of a spreadsheet summarizing the mapping of spectra to peptides. The total time from shipment of samples to receiving results was around 5 days.

The workflow on Platinum—including sample preparation, sequencing, and data analysis—was completed by one scientist. Proteins were digested into peptides and prepared for sequencing in less than 3 days and required less than 3 hours of hands-on time. Protein sequencing on the Platinum was initiated immediately following sample preparation and data collected during the 10-hour sequencing run was automatically and securely transferred to the Cloud for analysis. Peptides were automatically identified by the analysis software based on the fluorescence and kinetic properties of dye-labeled NAA recognizers as described previously.[2] The total time from protein digestion to results interpretation was less than 4 days.

While different peptides were identified via MS and Platinum, the CDNF protein was correctly identified in both approaches. MS analysis of CDNF resulted in the identification of 25 total peptides mapping to CDNF. This set of peptides included a number of peptides derived from incomplete Lys-C cleavage (Figure 2A).

In MS, the software commonly uses a search engine to sieve through acquired mass spectra to identify individual peptides by comparing the observed mass-to-charge ratios (m/z) with those expected for known peptides in pre-selected databases. The software also performs statistical analysis to estimate the probability that each peptide identified is correct. Subsequently, MS software tools assign identified peptides to specific proteins that contain those peptides—here, an inference algorithm is used to group peptides together into proteins and assign a probability score for each protein identified.

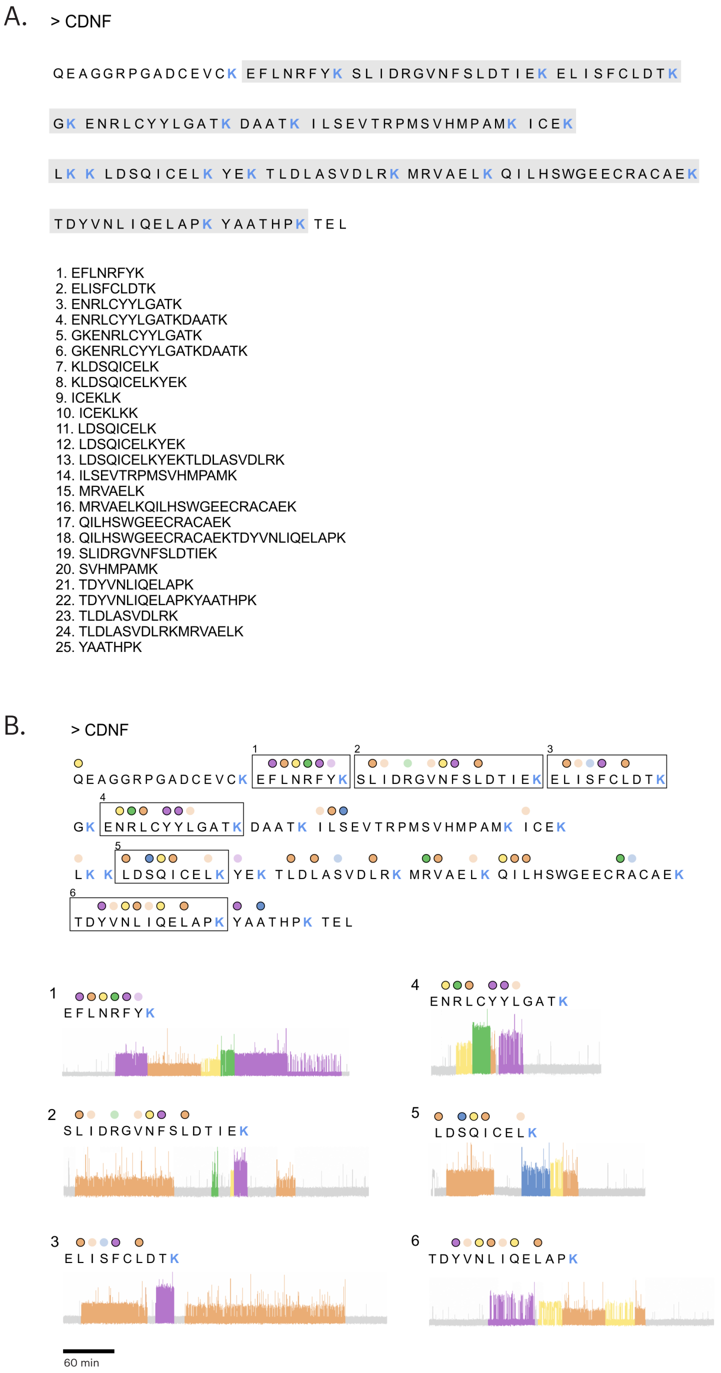

Figure 2. Comparison of Mass Spectrometry and Platinum Protein Analysis of CDNF: A.) CDNF Identification using Mass Spectrometry. B.) CDNF Identification using Platinum

Platinum analysis of CDNF resulted in unique identification of 6 peptides using kinetic binding signatures to correctly recognize amino acids and map them to the CDNF protein (Figure 2B). These peptides are sufficient to identify CDNF as the protein of origin from the human proteome with high confidence (Read the application note).

Both Platinum and MS resulted in identification of CDNF using distinct methods to confidently call unique peptides. The FGF2 and IL6 proteins were prepared and analyzed via Platinum and MS using the same workflow described for CDNF in Figure 1, and both resulted in correctly identifying FGF2 and IL6 proteins. However, the default MS analysis settings and software were unable to identify 1 of the 6 amino acid long peptides in FGF2 and a peptide containing chemically modified pyroglutamate in IL6 that were correctly identified by Platinum.

The full protein sequence of FGF2 has a total of 13 peptides, 8 identified by Platinum and 10 identified by MS (underlined and highlighted gray, respectively in Figure 3A). Similar to the CDNF dataset, both Platinum and MS correctly identified FGF2. However, the RLYCK peptide was not identified on MS because the default minimum peptide filter on MS was set to 6 amino acids in length to filter unwanted peptides. Platinum was able to identify the RLYCK peptide using its unique kinetic signature (Figure 3B).

A. PALPEDGGSGAFPPGHFKDPK **RLYCK** NGGFFLRIHPDGRVDGVREK SDPHIK LQLQAEERGVVSIK
GVCANRYLAMK EDGRLLASK CVTDECFFFERLESNNYNTYRSRK YTSWYVALK RTGQYK LGSK TGPGQK
AILFLPMSAK

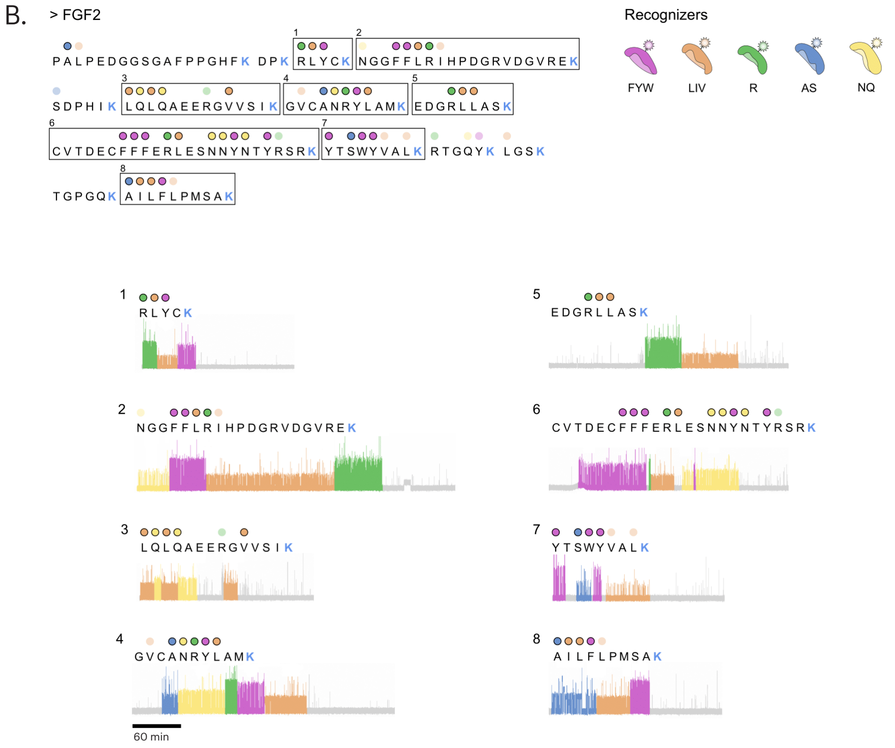

Figure 3. *Identification of FGF2: A.) Underlined peptides of FGF2 are identified by Platinum including the RLYCK peptide, and gray highlighted peptides of FGF2 are identified by Mass Spectrometry excluding the RLYCK peptide. B.) Kinetic Signatures of FGF2 on Platinum identify 8 peptides including RLYCK*

IL6 was also correctly identified by MS and Platinum. However, the QIRYILDGISALRK peptide was not detected with default MS analysis settings. The N-terminal glutamine residue of this peptide is chemically modified to pyroglutamate under certain conditions, such as sample preparation or LC-MS ionization. The original IL6 results obtained from MS did not report this peptide because a chemical modification to pyroglutamate would not be correctly detected using the default settings and database. Expert review of the results was required to enable detection of the modified pyroglutamate residue. Platinum results of IL6 were able to identify this peptide using its unique kinetic signature (Figure 4).

While both Platinum and MS can be used to correctly identify proteins, Platinum simplifies the workflow for investigating specific peptides or amino acids of interest without additional analysis.

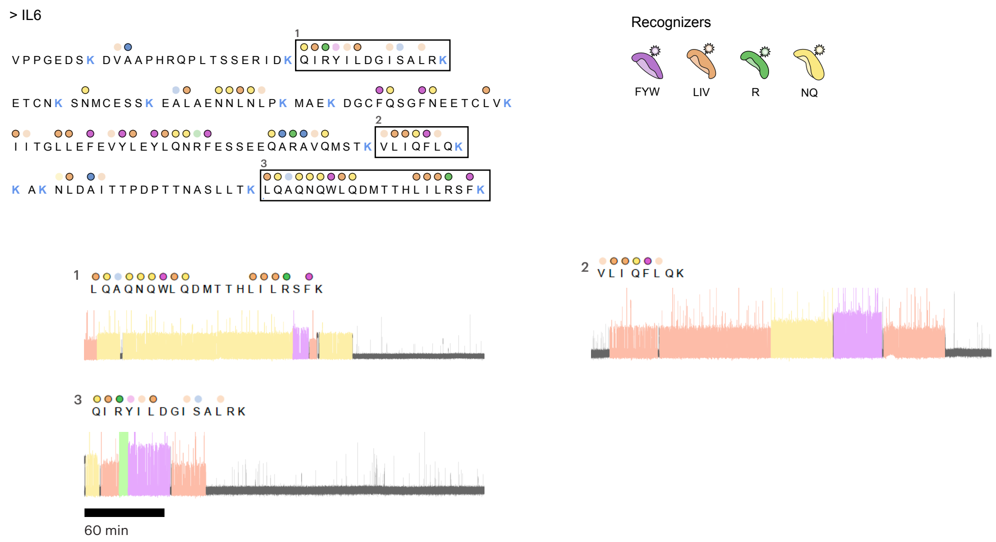

Figure 4. *Identification of IL6 on Platinum including QIR Peptide*

Conclusion

Quantum-Si's Platinum workflow offers a convenient solution for single-molecule protein identification without the need for expensive capital equipment and advanced expertise. Results are automatically provided in a Cloud-based software environment, enabling easy and straightforward interpretation of protein identification results. Single-molecule protein sequencing on Platinum offers a solution for deeper interrogation of proteoforms and PTMs through the use of kinetic signatures and enables researchers to make new discoveries about the proteome.

REFERENCES

1. Neagu AN, Jayathirtha M, Baxter E, Donnelly M, Petre BA, Darie CC. Applications of Tandem Mass Spectrometry (MS/MS) in Protein Analysis for Biomedical Research. *Molecules*. 2022;27(8):2411. doi:10.3390/molecules27082411

2. Reed BD, Meyer MJ, Abramzon V, et al. Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device. *Science*. 2022;378(6616):186-192. doi:10.1126/science.abo7651