QuantumSi

Protein Identification via Next-Generation Protein Sequencing and Proteome-Wide Mapping

SUMMARY

Protein identification is of fundamental importance in many areas of proteomics. Applications include determination of the presence or absence of an expected protein in a sample of interest, identification of an unknown protein present in a biological sample, and identification of a protein responsible for a biochemical activity in an isolated protein fraction. In some cases, mass spectrometry or affinity-based methods may be suitable options, but these methods can face substantial challenges when the protein of interest is unknown, with certain amino acid sequences or reliable detection of post-translational modifications (PTMs). To offer a solution based on direct sequencing and to provide more accessible tools for discovery in protein science, Quantum-Si has integrated our single-molecule sequencing output with automated cloud-based algorithms for proteome-wide mapping of sequencing data to enable accurate protein identification.

QUANTUM-SI TECHNOLOGY

Quantum-Si's Platinum™ instrument enables massively parallel peptide sequencing from biological samples. Our state-of-the-art semiconductor chip contains millions of wells that act as independent sequencing machines to achieve single molecule resolution. Time Domain Sequencing™ allows us to sequence individual peptides by measuring three dimensions of data—fluorescence intensity, lifetime, and binding kinetics—on an integrated semiconductor chip. This allows for the differentiation and identification of amino acid residues and PTMs. A more detailed overview of the Time Domain Sequencing™ and overall data analytics process can be found in our recently published White Paper.[2]

For Research Use Only. Not for use in diagnostic procedures.

2.0

Introduction

Linking proteolyzed peptides with their antecedent proteins is an ambitious goal in bottom-up proteomics, which relies on digesting intact proteins. In typical peptide-centric approaches, peptides are fractionated, and their spectra are matched with a database of simulated spectra generated via *in silico* protein digestion. However, several confounding factors limit the unambiguous, proteome-wide mapping of peptides. For example, missed or unanticipated cleavages and post-translational modifications can lead to peptides that are not detected by the search algorithm.[1]

Single-molecule protein sequencing offers an alternative method to identify proteins based on the kinetic signature of binding between recognizers and N-terminal amino acids. This approach provides the peptide-level resolution needed to discern peptides with similar sequences or physicochemical properties. To advance our technology for mapping to the proteome, we have developed cloud-based software that automatically maps sequencing data to the human proteome for protein identification.

3.0

Methodology & Workflow

Proteins are sequenced on Quantum-Si's Platinum instrument using our library prep and 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 N-terminal amino acids (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 determine amino acid sequence. We visualize the sequencing profiles of peptides 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).[3]

As described previously, when recognizers bind to NAAs, they also make important contacts with nearby downstream residues 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.

FIG. 1

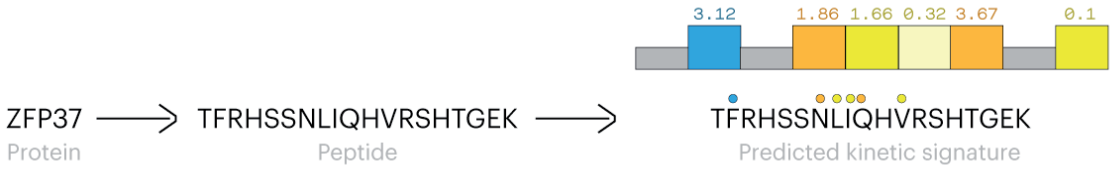

Fig 1. Example of kinetic signature prediction using Quantum-Si's kinetic model. A peptide fragment from in silico digestion of ZFP37, a zinc finger protein, was analyzed to predict its expected sequencing behavior in a real-time dynamic sequencing assay. Predicted RSs (colored boxes) are colored according to the corresponding binder; predicted average PD in seconds is indicated above each RS.

We developed a kinetic model that accurately predicts the PD for every possible 4-amino-acid sequence that starts with an N-terminal recognizer target. The kinetic model allows us to predict the kinetic signature for every peptide in a protein database of interest, for example the entire human proteome. An example of kinetic signature prediction is shown in Figure 1. We also developed analysis software that automatically identifies clusters of traces with highly similar patterns and generates an empirical kinetic signature for each cluster. With our kinetic model and clustering software, we can generate empirical kinetic signatures from protein sequencing data and pinpoint the protein of origin in the proteome by identifying peptides with matching predicted kinetic signatures.

For Research Use Only. Not for use in diagnostic procedures.

4.0

Results & Discussion

We used the human protein cerebral dopamine neurotrophic factor (CDNF, 161 amino acids) as a model protein to demonstrate protein identification from sequencing data based on our kinetic model and proteome mapping software.[4] We sequenced recombinant CDNF using a set of four recognizers, expanded from the set of three we previously demonstrated to include a new recognizer—PS1259—that recognizes N-terminal glutamine (Q) and asparagine (N) amino acids. This set of four recognizers recognizes a total of 9 NAAs (**Figure 2A)**.

We digested CDNF using the endopeptidase Lys-C and prepared a peptide library for on-chip sequencing. Sequencing analysis indicated that five peptides were expected to be readily observed on-chip because they are predicted to produce informative kinetic signatures with four or more RSs. Indeed, five main clusters of traces were identified in the sequencing output based on similarity of the pattern and kinetics of recognition using Quantum-Si's software analysis workflow (**Figure 2B**).

FIG. 2

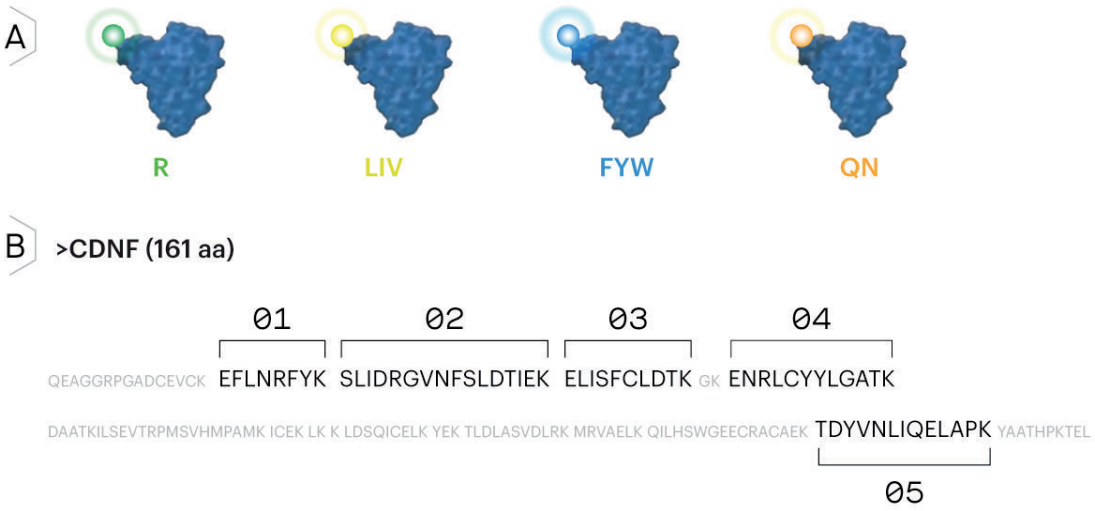

Fig 2. (A) Set of 4 NAA recognizers used to sequence CDNF protein in this study. B) The amino acid sequence of human CDNF protein with the 5 peptide fragments from Lys-C digestion that are expected to be readily observed on-chip indicated with bold lettering.

For Research Use Only. Not for use in diagnostic procedures.

Representative example traces for each cluster are displayed in **Figure 3**. The analysis software produced a characteristic kinetic signature summarizing the pattern of recognition and average PD for the traces grouped in each cluster. These kinetic signatures were then used as input into Quantum-Si's mapping algorithm to identify potential matches across the entire human proteome. The database of candidate peptides consisted of over 300,000 peptides of 8 or more amino acids in length derived from an *in silico* digest of the human proteome, representing roughly 20,000 human proteins.[5]

FIG.3

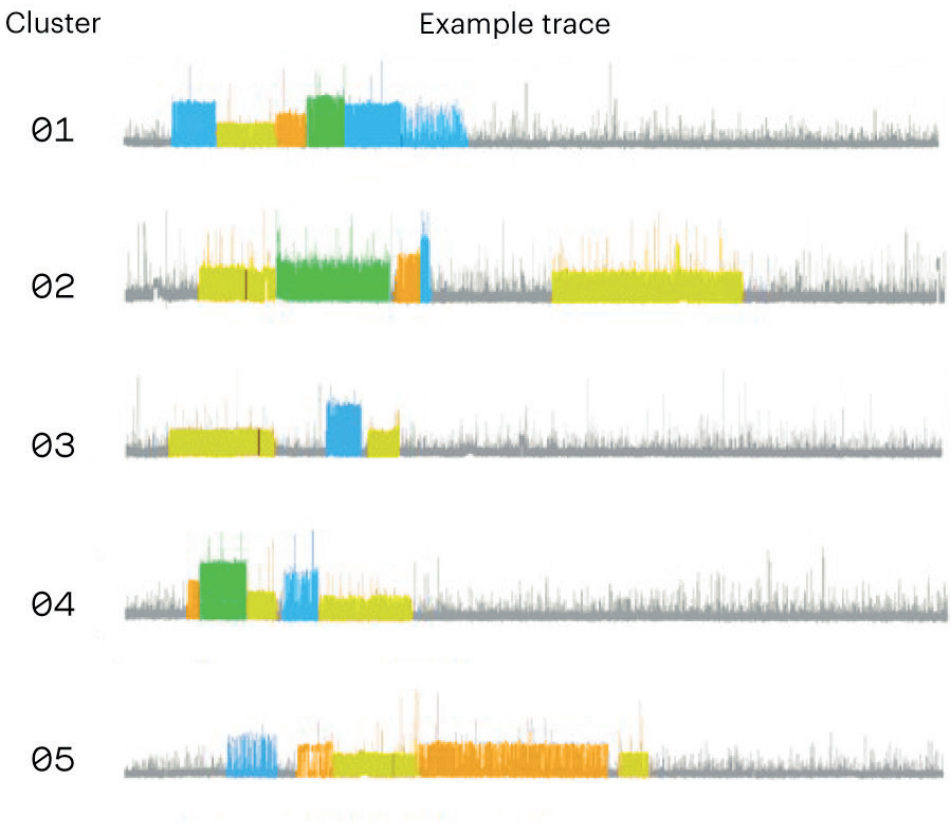

Fig 3. Example traces from each of the 5 independent clusters of traces generated in a sequencing run using recombinant human CDNF protein. Traces in each cluster were used to generate kinetic signatures for mapping to the human proteome.

AUTOMATED MAPPING OF SINGLE-MOLECULE SEQUENCING DATA TO THE HUMAN PROTEOME

FIG. 4

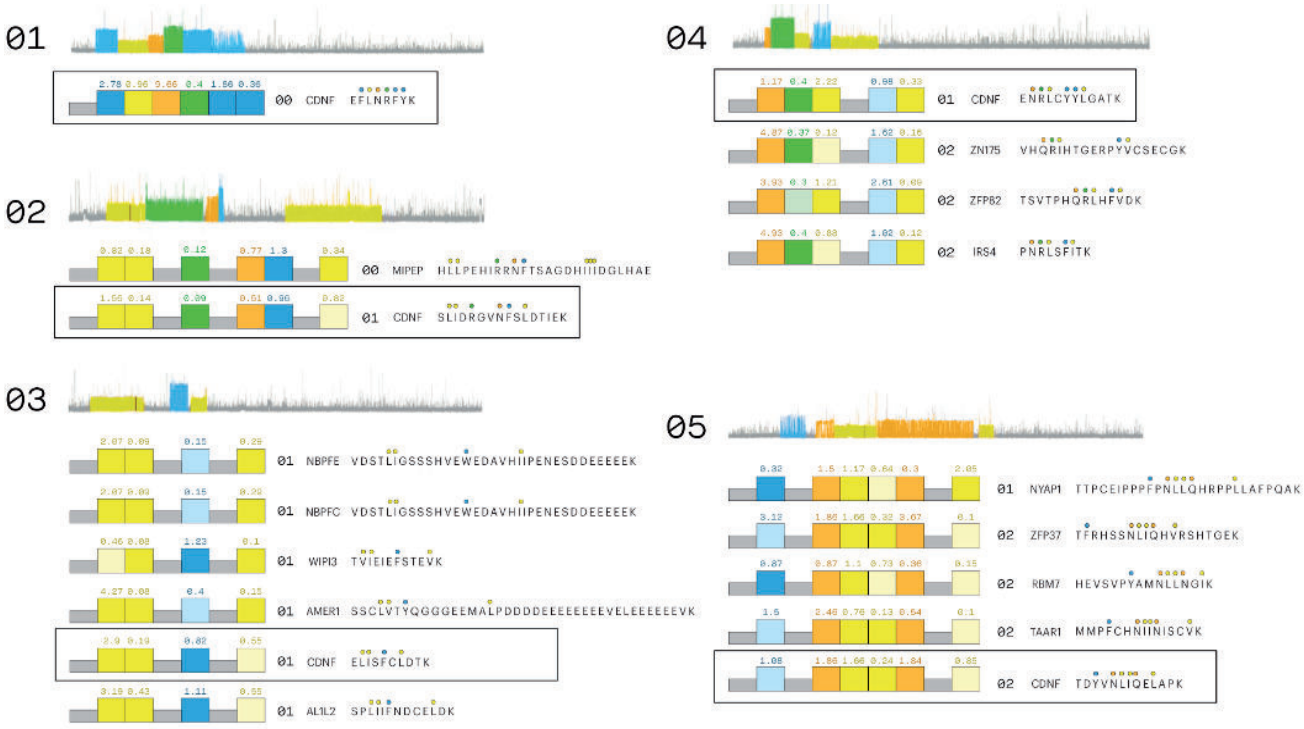

Fig 4. Mapping of kinetic signatures to the human proteome. Kinetic signatures from each of 5 clusters of traces from CDNF sequencing were mapped to a database of predicted kinetic signatures for >300,000 peptides derived from in silico Lys-C digestion of ~20,000 human proteins. Candidate matching peptides are shown for each cluster, with candidates corresponding to CDNF peptides outlined. The number to the right of each predicted kinetic signature indicates the number of RSs with predicted average PD that were not very close matches to the observed PD according to the kinetic model, useful as a ranking metric. Not all matches are shown for cluster 4; this cluster matched 229 total peptides due to its lower information content (4 RSs from only 2 recognizers).

For Research Use Only. Not for use in diagnostic procedures.

FIG. 5

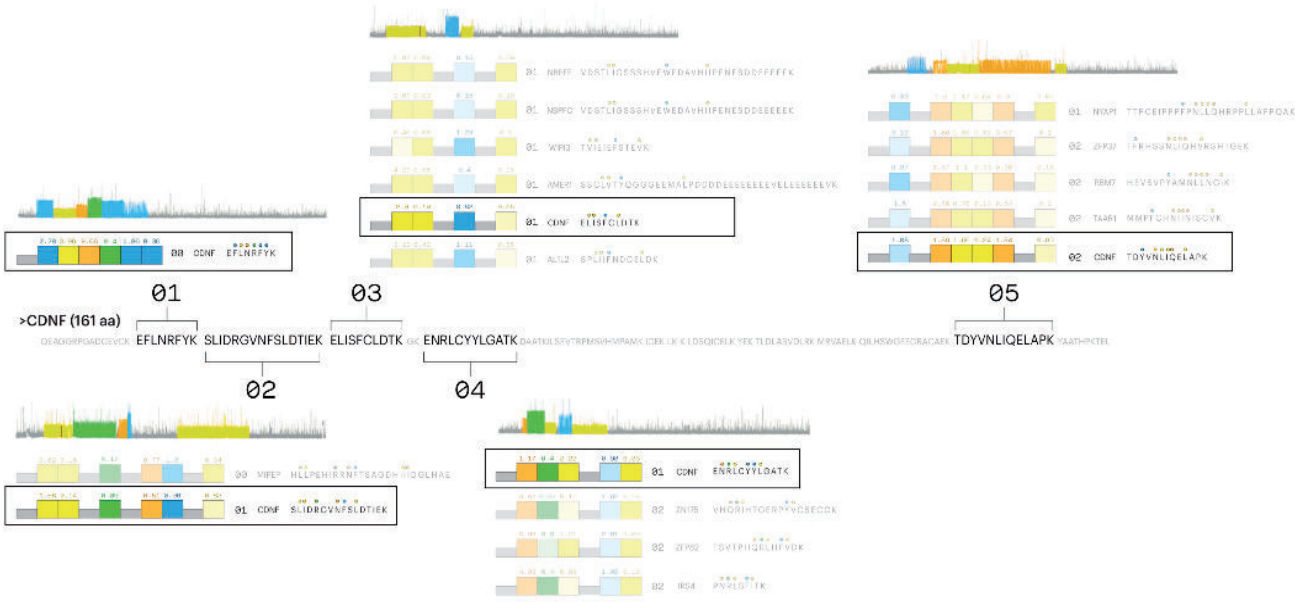

Fig 5. Matching peptides for each kinetic signature aligned to the full-length sequence of human CDNF protein.

The results of proteome-wide mapping are displayed in **Figure 4**. Each of the 5 kinetic signatures mapped to a set of candidate proteins that included CDNF as a top match or as the only match. Signatures containing 5 or more RSs were particularly successful at pinpointing CDNF, generating sets with 5 or fewer matching candidate proteins. Taken together, these results identified CDNF as the only human protein capable of generating the complete observed sequencing output with extremely high confidence (**Figure 5**).

For Research Use Only. Not for use in diagnostic procedures.

5.0

Conclusion

In this Application Note, we advanced our technology by matching sequencing data with the human proteome for protein identification. We show that sequencing data from multiple peptide fragments obtained from Quantum-Si's Platinum instrument can be used to generate highly characteristic kinetic signatures and identify a protein by mapping to the human proteome based on the predicted kinetic signatures of human peptides. To accomplish this task, we introduced a 4th recognizer and built a kinetic model and cloud-based software to automatically search the human proteome for protein identification.[6] The results presented demonstrate the capability of Time Domain Sequencing to identify known or unknown proteins in biological samples via proteome-wide mapping with high accuracy.

REFERENCES

1] Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat. Methods. 2019, 16, 509–518.

2] Digitizing Protein Sequencing: Detecting Individual Amino Acids through Real-Time Photon Emission. quantum-si.com. Quantum-Si. 2022.

3] Reed, B. D. et al. Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device. Science. 2022, 378, 186-192.

4] Lindholm, P.; Saarma, M. Cerebral dopamine neurotrophic factor protects and repairs dopamine neurons by novel mechanism. Mol Psychiatry. 2022, 27, 1310–1321.

5] Uhlén, M. et al. Tissue-based map of the human proteome. Science. 2015, 347, 1260419.

6] Alfaro, J. A. et al. The emerging landscape of single-molecule protein sequencing technologies. Nat. Methods. 2021, 18, 604–617.

For Research Use Only. Not for use in diagnostic procedures.