Shotgun Protein Sequencing With Meta-Contig Assembly

Adrian Guthals[1], Karl Clauser[3], Nuno Bandeira[1,2]

[1]Department of Computer Science and Engineering, University of California, San Diego, CA
[2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, CA
[3]Broad Institute of MIT and Harvard, Cambridge, MA

Introduction

Full-length de-novo sequencing of unknown proteins such as antibodies or proteins from organisms with unsequenced genomes remains an open problem. Conventional de-novo methods sequence MS/MS spectra individually, which yields poor accuracy and limited sequence length. Given these limitations, current techniques for sequencing unknown proteins rely on hybrid approaches involving de-novo sequencing followed by error-tolerant database search and/or homologous mapping to reconstruct protein sequences [1,2]. In contrast with current approaches, our approach aggregates and jointly sequences multiple spectra from overlapping peptides. Our Meta Shotgun Protein Sequencing (Meta-SPS) approach assembled unidentified MS/MS spectra into “meta” de-novo sequences up to 97 amino acids in length without any sequence homology steps and while mis-predicting only 1 in 33 amino acids.

Method

Tandem mass (MS/MS) spectra are de-isotoped by converting each isotopic envelope to its monoisotopic peak. SPS [3] is then used to sequence spectra from overlapping peptides into ‘contig spectra’, which have 18X less noise than input MS/MS spectra. Contigs are first aligned to one another to find where they can be reliably extended and then merged into longer ‘meta-contigs’ using a new spectrum alignment score and a new iterative assembly algorithm. Briefly, multiple contigs aligned to each other are joined by re-sequencing into a single 'meta-contig', which imposes new alignment scores with overlapping contigs that are then used to further extend the meta-contig.

Preliminary Data

To guarantee appropriate sequence coverage by acquired MS/MS spectra we analyzed a mixture of 6 known proteins. Spectra were acquired by LC-MS/MS using a Thermo LTQ Orbitrap to record both MS and MS/MS spectra at high resolution in the Orbitrap. The protein mixture was separately digested with different enzymes (trypsin, chymotrypsin, Lys-C, Glu-C, Asp-N, and Arg-C) and cyanogen bromide to yield a distribution of overlapping peptides of a variety of lengths. MS/MS spectra were acquired in both HCD and CID fragmentation modes. HCD spectra gave more complete b/y ion ladders on average than CID spectra. The majority of identified CID spectra had at least 40% of expected b/y ions, while the majority of identified HCD spectra had at least 60% of expected b/y ions. Because each fragmentation mode yields characteristic peptide fragmentation and noise distributions, spectra were de-isotoped and separately assembled into 463 CID and 763 HCD contigs. There were a total of 1226 SPS contigs of average length 9.5 amino acids, with the longest contig covering 36 amino acids. Contigs covered 88% of all targeted proteins with 93% accuracy, while collectively overlapping more than 6X the covered regions. Using this high degree of overlap, SPS contigs were ultimately assembled into 45 meta-contigs, each combining 4 contigs or more and covering 30 amino acids on average (3x longer than SPS contigs). Twelve of them covered 40 amino acids or more, with one of them covering 97 amino acids. Meta-contigs that combined 2 SPS contigs or more still covered 80% of target proteins at lower coverage redundancy of 1.2X in covered regions. Alignments were ultimately selected with 96% accuracy and 94% recall of alignments with 6 or more matching peaks. Only 1 in 33 amino acids of meta-contigs were called incorrectly (97% accuracy).

Novel Aspect

Approaching full length de-novo protein sequencing using new contig spectral alignment and iterative contig assembly techniques.

References

1. Castellana NE, Pham V, Arnott D, Lill JR, and Bafna V. Template proteogenomics: sequencing whole proteins using an imperfect database. Mol Cell Proteomics, 9:1260–70, Feb 2010.

2. Liu X, Han Y, Yuen D, and Ma B. Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. Bioinformatics, 25:2174–80, Sep 2009.

3. Bandeira N, Clauser KR, and Pevzner P. Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. Mol Cell Proteomics, 6:1123–34, Jul 2007.