Whole Protein de novo Sequencing from MS/MS

Overview
We present a semi-automatic workflow for characterizing antibody primary structure utilizing protein contigs assembled from de novo peptide sequences.

Introduction
De novo peptide sequencing is a practical technique for characterizing uncharted proteins. Among all proteins, antibodies are the most common subject to be sequenced. In general, it is required to achieve complete sequence coverage with multiple enzymatic digestion and produce evidence for every residue, specifically for the hypervariable CDR regions. While automated de novo peptide sequencing is routinely conducted, it is still puzzling and tedious to infer the original protein sequence from those short peptides.

Method
1. Assembling of protein contigs
A program is developed using a greedy algorithm for assembling protein contigs. It first detects the overlapping de novo peptides and iteratively extends the contigs using the consensus residues at both direction. In the extending process, contigs are merged if their sequences collide with each other. Protein contigs are extended until no more residues can be confidently added.

2. Workflow for antibody sequencing
   1. Assemble an antibody template
      - Blast protein contigs in NCBI nr database
      - Select a protein hit corresponding to the constant region
      - Select the closest protein hit corresponding to the variable region.
   2. Draft a sequence using the contigs and template
      - Map the contigs to the antibody template
      - Solve disagreements between contigs and the template
        - Correct typical de novo sequencing errors (I/L, AG/Q, PK/KP, variable PTMs)
        - In principle, trust contigs in variable region and the template in constant region
      - Solve the residues not covered by contigs, if any, using de novo peptides
   3. Refine the sequence using PEAKS’ SPIDER homology search
      - Perform SPIDER homology search on the draft sequence
      - Examine insertion/deletion/mutation reported by SPIDER
      - Examine residues with low peptide coverage
      - Examine residues at the protein n-terminus
      - Compare the sequence mass with protein intact mass, if available
      - Make corrections and repeat step 3 until the sequences converge.

Result
1. Experiment
The antibody, I5154 (human IgG1/kappa), was purchased from Sigma-Aldrich to evaluate the proposed workflow. The sample was reduced with DTT, alkylated with iodoacetamide. Glycans were removed and heavy/light chains were separated. Each chain was digested with six enzymes AspN, chymotrypsin, GluC, LysC, pepsin and trypsin. MS/MS spectra was acquired using LTQ-Orbitrap at high resolution with HCD fragmentation.

2. Blind trials
The proposed workflow was also tested in a series of blind trials. Our collaborators provide antibody mass spec data to us for in-house analysis, while not disclosing the true sequences. The sequencing results were correct at a minimal of 96% of residues, meanwhile all the wrong residues were previously identified with low confidence during the sequencing analysis.

Conclusion
Protein sequencing can be efficiently conducted by utilizing protein contigs assembled from de novo peptides. The described semi-automated workflow provides a reliable solution for antibody sequencing with MS/MS of multiple enzymatic digestion. The proposed workflow has been routinely used for data analysis in the CHAMPS - Antbody Sequencing Service, provided by Bioinformatics Solution Inc. (http://www.bioinfor.com/peaks/products/champs.html)