Novor and REMAb
De Novo Peptide Sequencing Accuracy

- For below average spectrum, as much as 50% error rate!
- Error source:
  - spectrum quality
  - Inaccurate scoring function
- Solutions:
  - Improve instruments.
  - Improve scoring
  - Make use of partially correct sequences.
Amino Acid Combination Affects the Peak Intensity

- The neighbouring 3 amino acids approximately determine the peak intensity.
An Example: Proline (P)

- Most software’s scoring function prefers more abundant peaks.
- Pro enhances fragmentation at left, and reduces at right.
Scoring Features

- $\text{Pr}(E \text{ is correct})$ predicted by features such as
  - mass error
  - intensity of $y_6$ and $y_7$
  - intensity ratio $y_6/y_7$
  - $L$, $E$, and $N$
  - and many others
Decision Tree

- Somewhat Confident
- Very Confident

Not Confident
- Yes: Are fragment ions abundant at left side?
  - No
- Yes: Are fragment ions observed at both sides?
  - No
- Yes
Decision Tree Learning

NIST Spectrum Library
340,000 spectra

169 features
14,000 internal nodes
average depth 18.4
Benefits of Decision Tree

• Allows to use of a large number of scoring features
  • Mass error, sequence pattern, all ion types, intensity, etc.
  • 169 features

• Learn a large number of rules
  • 14,000 branching nodes

• Each evaluation is fast.
  • Path from root to a leaf average length = 18
  • Only most important features are examined according to situation.
Algorithm

• A peptide’s score is the sum of amino acid confidence score.
• Algorithm computes a peptide to maximize this score.
Novor vs. PEAKS (Accuracy)

Precision vs. Recall

Precision = \frac{\# \text{ Correct AA above threshold}}{\# \text{ AA above threshold}}

Recall = \frac{\# \text{ Correct AA above threshold}}{\text{total \# AA}}

AA: Amino Acid
Novor vs. PEAKS (speed)

- Speed is an order of magnitude faster.
- First and only real-time de novo sequencing software.

De Novo Protein Sequencing Basic Idea

1. Digest protein with different enzymes.
2. De novo sequence each peptide.
3. Assemble overlapping peptides to derive the protein sequence.
Antibodies

- Antibody-antigen specific binding.
- Over 2 million antibody products, wildly used in research, diagnosis, and drugs.
- Global sales of antibody drugs over 100 billion$ annually.
- Human serum antibodies reflect the immune system activities.
REmAb™ Project

• Goal: Automated routine mAb protein sequencing with mass spec.
• Joint effort with Rapid Novor Inc. since 2016.
• Approach:
  • Manually sequence hundreds of antibodies. (250+ sequenced to date).
  • Gradually improve the level of automation with software.
  • Optimize the mass spec and algorithm simultaneously.
  • Statistical validation of the sequencing result.
  • Commercialization for sustainability.