**1. Overview/Introduction**

**Motivation**

On several occasions, proteins from different species have originated useful drugs in combating human conditions such as:

- Severe chronic pain associated with cancer/AIDS (Prolix by Genentech)
- Hypertension (Capoten by Bristol-Myers Squibb)
- Type-2 diabetes (Avandia by Aventis)
- Blood clotting (Integrilin by Millennium)

Many of these proteins came from organisms whose DNA sequence is unknown like scorpions and different snake species.

**Problem**

Current methods for de novo sequencing of whole proteins are labor intensive and very restrictive (Edman microsequencing).

**Contribution**

We propose a fully automated high-throughput Shotgun Protein Sequencing approach to sequence mixtures of modified proteins.

---

**2. Methods: Standard input data**

Protein mixture

Set of overlapping peptides

Protein sequence

Set of MS/MS spectra

---

**2.1 Motivation**

Set of overlapping peptides

Protein mixture

---

**3. Methods: High-throughput protein sequencing**

**Stage 1: Separation of b/y ions by spectral alignment**

Spectra are preprocessed to be fully symmetric (peak masses and intensities)

The diagonal diagonal is the alignment of b/y ions in both aligned spectra, respectively.

**Stage 2: Merge matched b-ions**

Note that input spectra now contain almost only b ions.

Every spectrum is converted to a spectrum graph.

- Each peak corresponds to a vertex.
- Vertices are connected by an edge if the corresponding peaks differ by an amino acid mass.

**Stage 3: De-novo sequencing**

The set of aligned spectra is converted to a single spectrum graph.

- The simplest objective function would simply select the path that explains the largest number of edges in the graph. A maximum likelihood solution can be derived from each vertex’s support to each aligned spectrum.

---

**4. Experimental results**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Extensive sequence coverage</th>
<th>Accurate de novo sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA_ATROX</td>
<td>92%</td>
<td>90%</td>
</tr>
<tr>
<td>VT_A</td>
<td>91%</td>
<td>85%</td>
</tr>
<tr>
<td>MM</td>
<td>94%</td>
<td>90%</td>
</tr>
<tr>
<td>MM</td>
<td>92%</td>
<td>91%</td>
</tr>
</tbody>
</table>

---

**5. Conclusions**

**Accurate:** amino acid prediction accuracy of 92% on the test sample (10Kb protein, several contaminants)

**Robust:** amino acid prediction accuracy remains very high (90%) on a mixture of snake venom proteins

**Extensive:** recovers almost all the amino acid content identified by database search

**Reproducible:** similar protein reconstruction results achieved for different mixtures of modified proteins