**CFM-ID: A Web Server for Annotation, Spectrum Prediction and Metabolite Identification from MS/MS**

**Felicity Allen, Allison Pon, Michael Wilson, Russ Greiner, David Wishart**
Department of Computing Science, University of Alberta

Contact: felicity.allen@ualberta.ca

---

**Summary**

- **Goal:** Automated identification of metabolites from tandem mass spectra (MS/MS).
- **Existing Methods:**
  - Search against reference databases of measured spectra.
  - Enumerate all ways molecules could break and/or make a heuristic selection of likely breaks.
  - Predict spectra—usually predict far more peaks than actually occur.
- **Approach:**
  - Design Competitive Fragmentation Modeling (CFM), a model for Electrospray (ESI) MS/MS fragmentation.
  - Derive parameters for CFM from MS/MS data.
- **CFM-ID:**
  - A web server that uses CFM to provide three utilities associated with interpretation of MS/MS spectra:
    - Spectrum Prediction, Peak Assignment and Compound Identification.
- **Experimental Results:**
  - Spectrum Prediction: Better Jaccard scores vs full enumeration of possible peaks.

---

**Competitive Fragmentation Modeling (CFM)**

- Model ESI-MS/MS (above) fragmentation as a stochastic, homogenous, Markov process of state transitions between charged fragments (below).

**Single Energy (SE-CFM)**

- Observation model links $F_i$ to $P$ via Gaussian distribution.

**Combined Energy (CE-CFM)**

- The initial molecule ($F_i$) and the output peak ($P$) are observed.
- All intermediate fragments ($F_j$) are latent.
- Possible transitions: Enumerate a graph of all possible fragmentations for each molecule (right), similar to $\lambda^{n!}$.
- Softmax transition function is competitive: a particular break is likely to occur only if no other breaks are substantially more likely.

- Given $\Phi_i$, chemical features associated with break $(f_i, f_j)$, assign $0/\Phi_i$: $\phi_i$.

- Break Pair:
  - C-C bond, C-N bond, etc.
- Root paths (length 2 and 3):
  - C-N on side? true
- Gasteiger Charges of root atoms
- Hydrogen Movement
- 1H moves from N to ion
- Ring Break Features
- Size of ring, broken bond distance, aromatic...

**Experimental Validation**

**Data Sets:**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Mols</th>
<th>Mode</th>
<th>Device</th>
<th>Energies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metlin (+)</td>
<td>1409</td>
<td>+</td>
<td>Agilent 6510 Q-ToF</td>
<td>10V, 20V, 40V</td>
</tr>
<tr>
<td>MassBank</td>
<td>192</td>
<td>+</td>
<td>Agilent 6520 Q-ToF</td>
<td>10V, 20V, 40V</td>
</tr>
<tr>
<td>HMDB</td>
<td>500</td>
<td>+</td>
<td>Waters Quattro</td>
<td>10V, 25V, 40V</td>
</tr>
<tr>
<td>Metlin (-)</td>
<td>976</td>
<td>-</td>
<td>Agilent 6510 Q-ToF</td>
<td>10V, 20V, 40V</td>
</tr>
</tbody>
</table>

**Spectrum Prediction:**

- Compare vs full enumeration of all possible fragments (right).
- Low energy (10V) spectra better predicted.
- Positive though imperfect correlation between measured and predicted intensities values—Pearson correlations of 0.7 (10V), 0.6 (20V) and 0.45 (40V).

**Compound Identification:**

- Query KEGG and PubChem for candidates within tolerance of the known mass of the target.
- Compare against other methods (below).

**CFM-ID Web Server**

- Supports three sub-tasks for automated metabolite identification from MS/MS data:
  - **Spectrum Prediction:**
    - Runs trained CFM model forward to predict spectra for low, medium and high collision energies.
  - **Peak Assignment:**
    - Assigns fragments within mass tolerance of each peak.
  - **Compound Identification:**
    - Predicts spectra for all candidate compounds.
  - **Experimental Validation:**
    - Ranks compounds by Jaccard score before measured and predicted spectra.

---

**References**