iPRG 2011:


ABRF 2011, San Antonio, TX
2/20/11
INTRODUCTION:
CID AND ETD FROM 30K FEET
Collision Induced Dissociation (CID) relies on a series of bimolecular events (collisions) to provide the peptide precursor with sufficient energy to fragment (ergodic process). CID typically causes backbone fragmentation. y and b ions are by far the most prevalent fragment types.

Electron Transfer Dissociation (ETD) relies on the transfer of a single electron to a peptide precursor. This transfer likely creates a radical that very quickly decays into ion fragments (a non-ergodic process). Like CID, ETD typically causes backbone fragmentation, but mostly resulting in c and z ions.
CID and ETD spectra - example

NanoLC-ESI-MS/MS analysis of the CcD subunit IV.

CID

ETD

Hefting S et al. Mol Cell Proteomics 2009;7:1714-1724

©2009 by American Society for Biochemistry and Molecular Biology
iPRG 2011 STUDY: CONCEPT
Study Goals

- **Primary:** Evaluate the ability of participants to identify ETD spectra
- **Secondary:** Find out why result sets might differ between participants
- **Tertiary:** Produce a benchmark dataset, along with a spectral library and an analysis resource
Study Design

- Use a common, rich dataset
- Use a common sequence database
- Allow participants to use the bioinformatic tools and methods of their choosing
- Use a common reporting template
- Report results at an estimated 1% FDR (at the spectrum level)
- Ignore modification localization
- Ignore protein inference
The sample

*NIST yeast lysate (six vials of RM8323), 228μg protein, LysC digest separated on SCX column*

Sample prep by
Robert Chalkley, 
UCSF

SCX by
Jinal Patel, 
The Broad Institute
Choosing a fraction for the study

Precursors selected for CID / ETD

QAD identifications from CID / ETD
Study Materials (i)

- 1 LTQ-Orbitrap XL dataset (eq. 1 RAW file)
  - RAW, mzML, mzXML, MGF, dta – conversions by ProteoWizard 2.1.2051
- 1 fasta file (UniProtKB/SwissProt S. cerevisiae from Sept. 2010)
- 1 spectral library in SpectraST format (contributed by Henry Lam)
- 1 template (Excel)
- 1 on-line survey (Survey Monkey)
### Study Materials (ii) – additional data

<table>
<thead>
<tr>
<th>Frac</th>
<th>Instrument</th>
<th>Fragmentation</th>
<th>MS/MS Res/Acc</th>
<th>Spike?</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>D100914_yeast_SCK09_rak_110IT_pc_01.RAW</td>
<td>XL</td>
<td>OT (C or E)</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>K100923_Yeast_SCK09_116DT_pcc_01.raw</td>
<td>V</td>
<td>C+E</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>K100923_Yeast_SCK09_1666E_pcc_01.raw</td>
<td>V</td>
<td>H+E</td>
<td>HH</td>
</tr>
<tr>
<td></td>
<td>D100915_yeast_SCK05_rak_11E_pc_02.RAW</td>
<td>XL</td>
<td>E</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td>V20100923-23</td>
<td>V</td>
<td>C+E</td>
<td>LL</td>
</tr>
<tr>
<td>10</td>
<td>D100914_yeast_SCK10_rak_110IT_pc_01.RAW</td>
<td>XL</td>
<td>OT (C or E)</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>K100923_Yeast_SCK10_116DT_pcc_01.raw</td>
<td>V</td>
<td>C+E</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>K100923_Yeast_SCK10_1666E_pcc_01.raw</td>
<td>V</td>
<td>H+E</td>
<td>HH</td>
</tr>
<tr>
<td></td>
<td>D100917_yeast_SCK10_rak_11E_pc_01.RAW</td>
<td>XL</td>
<td>E</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td>D100910_yeast_SCK10_rak_11E_pc_01.RAW</td>
<td>XL</td>
<td>E</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td>V20100923-24</td>
<td>V</td>
<td>C+E</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>V20100923-29</td>
<td>V</td>
<td>OT (C or E)</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>V20100923-31</td>
<td>V</td>
<td>E</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td>V20100923-32</td>
<td>V</td>
<td>OT (H or E)</td>
<td>HH</td>
</tr>
<tr>
<td></td>
<td>V20100923-32</td>
<td>V</td>
<td>OT (H or E)</td>
<td>HH</td>
</tr>
<tr>
<td></td>
<td>K100923_Yeast_SCK11_116DT_pcc_01.raw</td>
<td>V</td>
<td>OT (H or E)</td>
<td>HH</td>
</tr>
<tr>
<td></td>
<td>K100923_Yeast_SCK11_1666E_pcc_01.raw</td>
<td>V</td>
<td>H+E</td>
<td>HH</td>
</tr>
<tr>
<td></td>
<td>D100917_yeast_SCK11_rak_11E_pc_02.RAW</td>
<td>XL</td>
<td>E</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td>V20100923-25</td>
<td>V</td>
<td>C+E</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>V20100923-30</td>
<td>V</td>
<td>OT (C or E)</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>V20100923-33</td>
<td>V</td>
<td>OT (H or E)</td>
<td>HH</td>
</tr>
<tr>
<td>12</td>
<td>D100914_yeast_SCK12_rak_110IT_pc_01.RAW</td>
<td>XL</td>
<td>OT (C or E)</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>K100923_Yeast_SCK12_116DT_pcc_01.raw</td>
<td>V</td>
<td>C+E</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>K100923_Yeast_SCK12_1666E_pcc_01.raw</td>
<td>V</td>
<td>H+E</td>
<td>HH</td>
</tr>
<tr>
<td></td>
<td>V20100923-26</td>
<td>V</td>
<td>C+E</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>V20100923-27</td>
<td>V</td>
<td>E</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td>V20100923-28</td>
<td>V</td>
<td>E</td>
<td>LL</td>
</tr>
</tbody>
</table>

**Frac** – Fragment

**Instrument**
- XL – Orbitrap XL
- V – Velos Orbitrap

**Fragmentation**
- DT – Decision tree CID or ETD
- OT – Decision tree HCD or ETD
- C+E – CID, ETD on each precursor
- H+E – HCD, ETD on each precursor

**Spike?**
- LL – Low
- HH – High

---

**Not public yet**

Used to create library
Instructions to Participants

1. Retrieve and analyze the data file in the format of your choosing, with the method(s) of your choosing

2. Report the peptide to spectrum matches in the provided template

3. Fill out the survey

4. Attach a 1-2 page description of the methodology employed
# ABRF IPRG 2011 Study Template: ETD Data Analysis

**Instructions:** Please fill in all fields required fields (marked with ‘*’). After deleting the example rows, create a new row for each peptide spectrum match. Indicate whether each match is better than a 1% FDR on the spectrum-level. Include identifications above and below threshold. Results should be sorted by Search Engine Score from most to least confident. Additional instructions can be found above each field header. Results should be emailed to ‘anonymouse.iprg2011@gmail.com’ no later than Dec. 10, 2010. Please make sure to fill out the REQUIRED survey (URL).

<table>
<thead>
<tr>
<th>Spectrum Identifier**</th>
<th>Mass Error (ppm)</th>
<th>Charge*</th>
<th>Peptide Sequence*</th>
<th>Modifications</th>
<th>Protein Accession(s)</th>
<th>Search Engine Score</th>
<th>Better than 1% FDR threshold(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scan 1753</td>
<td>56.783</td>
<td>+4</td>
<td>PPPGSLQYK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scan 4669</td>
<td>86.429</td>
<td>+6</td>
<td>PRPPFRESWRRAQK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scan 2326</td>
<td>67.21224</td>
<td>+4</td>
<td>PSMDRQDIPNFAEKP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scan 3271</td>
<td>61.8987</td>
<td>+6</td>
<td>NITROQYK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scan 4022</td>
<td>53.1457</td>
<td>-1</td>
<td>2QSVLWYQG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scan 4732</td>
<td>67.1325</td>
<td>+6</td>
<td>2YHNNPDQYK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scan 5857</td>
<td>65.1482</td>
<td>+6</td>
<td>14AADLKD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Scan 7213             | 91.7753          | +6     | 4GLVDDMDAS3LNIQYVDFVNL3EACARBON4athyl | P54839 | 0.999999 Y

*Identifiers should be unique across all datasets.*

**Spectrum identifiers are unique across all datasets."
iPRG 2011 STUDY:
PARTICIPATION
Soliciting Participants and Logistics

Study advertised on the ABRF website and listserv, Molecular and Cellular Proteomics blogsite, ECD/ETD conference attendants, GenomeWeb and by direct invitation from iPRG members

1. Email participation request to ‘iPRGxxx@gmail.com’
2. Send official study letter with instructions
3. All further communication (e.g., questions, submission) through ‘iPRGxxx.anonymous@gmail.com’

“Anonymizer”
Participants (i) – overall numbers

- 40 requests / 36 submissions (‘90% return’)
  - Some participants submitted two result sets

- 9 initialed iPRG member submissions (with appended ‘i’)

- 8 vendor submissions (identifiable by appended ‘v’)
Participants (iii) – study opinions

Study Difficulty Level
- Easy: 51%
- Just right: 29%
- Challenging: 20%

Reporting Difficulty Level
- Easy: 35%
- Just right: 38%
- Challenging: 27%

Have you participated in previous ABRF studies?
- Yes: 38%
- No: 62%

Do you provide this service?
- Not Applicable: 12.27%
- No: 21.46%
- No but have plans to do so: 4.99%
- Yes: 18.18%
Participants (iv) – methods (i)

![Data Format Used](chart)

<table>
<thead>
<tr>
<th>Data Format</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGF</td>
<td>25</td>
</tr>
<tr>
<td>RAW</td>
<td>25</td>
</tr>
<tr>
<td>mzXML</td>
<td>10</td>
</tr>
<tr>
<td>DTA</td>
<td>2</td>
</tr>
<tr>
<td>mzML</td>
<td>2</td>
</tr>
</tbody>
</table>
Participants (v) – methods (ii)

Spectral Pre-Processing

Other / in-house (public / non-public):
DTARefinery, DeconMSn, DTA Generator, Etdgenerator, RawExtractor, Hardklor, multiplierz, ReAdW, Byonic
Participants (vi) – methods (iii)

Peptide identification

Other / in-house (public / non-public):
pFind, ByOnic, ProteinScape, MS_LIMS, PVIEW, PepArML, Byonic2, Proteome Discoverer
Participants (vii) – methods (iv)

Results filtering

Other / in-house (public / non-public):
pBuild, ComByne, ProteinScape, Percolator, PVIEW, Epitomize, FDR Optimizer, MSblender, OmssaParser, MascotDatFile, multiplierz, ComputeFDR, Proteome discoverer
Participants (viii) – *time spent (hours)*

- Spectral pre-processing
- Reformatting FASTA DB
- Search time
- Filtering for 1% FDR
- Manual validation
- Preparing Excel template
- Total analysis time
Participants (ix) – confidence

Confidence in processing ETD data

Before study
After study

No experience, never tried it before
Not confident
Confident
Very confident
iPRG 2011 STUDY:
PRELIMINARY ANALYSIS
iPRG studies are not competitions. Leftmost is not meant to imply best, it just reflects the sorting criterion: total number of confident ids; this was chosen as a convenient means of sorting, and this sort is used throughout for consistency.
Questions marks and blanks are present when participants either did not clearly indicate which modifications were allowed for or did not include them in a sequence specific manner in their results.
Since an inordinate number of spectra had only 1 or 2 participants agreeing, we selected 3 participants agreeing as the threshold for denoting consensus agreement. Consensus requires agreement on sequence, so do note that this still allows for disagreement on modification localization.
We would also liked to have categorized how many identified spectra were derived from enzymatic specificity of full, semi, or none. This information was not readily collected. However, for some participants 10-20% of confident identifications came from something other than full enzymatic specificity.
The green (NS) bars represent the room for improving confidence threshold setting. If one could improve the decision making about confidence in a peptide spectral match, without increasing FDR then the sum of the heights of the blue (YS) and green (NS) bars appears to be within reach for many participants to substantially improve their overall identification totals.
When particular identifications are reported by less than 3 participants it is difficult to tell whether that represents extraordinary skill or just another false positive. On the other hand, disagreement with the consensus (YD) is more likely to indicate a wrong answer. Consequently, the YD rate serves as a surrogate for the minimum FDR level. Note that many participants, especially those to the far left tend to have a YD rate >1%, which suggests they have underestimated their FDR level. The study was requested to be performed at 1% FDR.
Complete spreadsheet will be available for download at http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm
iPRG 2011 STUDY:
TWO SURPRISES AT THE END
Sigma 48 digested separately and not subjected to SCX. Spike-in at a level to yield ~10% of ID peptides. Biases against non-IPRG DB users, and SCX prediction users.
Sigma-48 as TP estimator

Proxy for FD rate

Sigma48 IDs - (Sigma48 IDs * YDR)
Identification of fusion peptides

Five participants reported the peptide:

KLVAASQAALGLMNYLETQLNKK

C-terminus of Human Serum Albumin - N-terminus of Pachytene arrest protein SAE3

Consensus Answer:

Acetyl-SRSGVAVADESLTAFNDLKLGK(Carboxymethyl)K²⁺

rare mod
Conclusions

- Study went well and had global participation, with quite a few first-timers joining in
- Earlier software problems with interpreting doubly-charged precursors have been largely cleared up
- Experience with software is probably a better measure of performance than the actual tool used
- People are generally over-optimistic about how reliable their results are (FDR underestimation)
- However, false negatives (NS) are generally much higher than false positives, so there is room for improvement there
What did the participants think?

"I work in an environment without a group of peers doing proteomics. So having a study like this definitely gives me a chance to compare notes with others and benchmark my abilities. It also gives me a piece of evidence to prove my ability to people that are not in the field, such as users, administrators, advisory committee. I believe all core only facilities should participate in these studies."

100% of participants found the study useful

"The results of this study will be good to show how much room for improvement there is for the popular identification tools in ETD analysis. It’s a good opportunity for lesser known and more open software to make a significant impact in the field."
THANK YOU TO ALL
STUDY PARTICIPANTS!

iPRG

Manor Askenazi
Nuno Bandeira
Robert Chalkley
Karl Clauser
Eric Deutsch
Henry Lam (ad-hoc member)
Paul Rudnick
Tom Neubert (EB liaison)
Hayes McDonald (chair elect)
Lennart Martens (chair)
John Cottrell (member elect)
Matt Chambers (member elect)
Ruixiang Sun (member elect)
Eugene Kapp (member elect)

Indispensable help came from:
Jinal Patel, The Broad Institute
Namrata Udeshi, The Broad Institute
Jeremy Carver, UCSD