Robust Method for Calculating the Local FDR for Database Search Results

ASMS 2010

Wilfred H. Tang, Agilent Technologies, Santa Clara, CA
Maintaining control over false positives is a significant challenge in analyzing proteomics data from tandem mass spectrometry experiments. The rate of occurrence of false positives in database search results is often assessed by calculating the false discovery rate (FDR).

Definitions

• Global FDR – Measures the FDR of a collection of IDs
• Local FDR – Measures the FDR of an individual ID

Currently, the global FDR is reported far more commonly than the local FDR, but there are some significant advantages to the local FDR.

Advantages of Local FDR

• If one is interested in a specific ID (for example, to pursue follow-up experiments), the local FDR is a more useful metric than the global FDR. The local FDR provides an estimate of the expected number of false positives for the specific ID of interest, while the global FDR only provides an estimate of the expected number of false positives for a collection of IDs (of which the ID of interest is a part of). In other words, the local FDR provides a more direct measure of the “pain ratio” – the cost of each additional correct ID in terms of incorrect IDs that lead to “wild-goose chases”
• Better metric for combining results from multiple iterations of database search

Disadvantages of Local FDR

• More difficult to calculate
• Higher error bars

We present here a simple yet general method for calculating the local FDR and demonstrate an implementation of this method in Spectrum Mill.
The key to successfully computing the local FDR is getting a good fit to \( F(N) \). We have tested our constrained least-squares fitting method on a variety of proteomics samples (of which 3 are shown here) and obtained decent fits in each case.

**Local vs. Global FDR**

The local and global FDRs can differ significantly. In this example, if one used an acceptance threshold of 5% global FDR, the error rate in the resulting set of IDs is 5% (1 in 20), but note that the error rate at the tail of the set is given by the local FDR and is 37% (more than 1 in 3).
Robust Method for Calculating the Local FDR for Database Search Results

Example of Spectrum Mill’s FDR analysis. Spectrum Mill reports both the global and local FDRs in tables and graphs and also shows the fit for $F(N)$. The analysis is performed at 2 different levels – the spectral level and the distinct peptide level.

### False Discovery Rate Search #1

#### Spectral Level Analysis

<table>
<thead>
<tr>
<th>Number of spectra detected</th>
<th>FDR Value</th>
<th>Local FDR</th>
<th>Global FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>4147</td>
<td>0.0857</td>
<td>0.1321</td>
</tr>
<tr>
<td>0%</td>
<td>2916</td>
<td>0.0804</td>
<td>0.1221</td>
</tr>
</tbody>
</table>

#### Distinct Peptide Level Analysis

<table>
<thead>
<tr>
<th>Number of distinct peptides detected</th>
<th>FDR Value</th>
<th>Local FDR</th>
<th>Global FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>12434</td>
<td>0.0684</td>
<td>0.1013</td>
</tr>
<tr>
<td>0%</td>
<td>1233</td>
<td>0.0638</td>
<td>0.0975</td>
</tr>
</tbody>
</table>

### Global and local false discovery rates

![Graph showing global and local FDR](image)

### Fit quality for computing local false discovery rate

![Graph showing fit quality](image)

Workflows (new in Spectrum Mill B.04.00) enable the user to easily specify multiple iterations of searching and to combine the results based on the local FDR.

Example workflow:
1. Identity mode search
2. Autovalidate at 5% local FDR
3. Variable modifications search on validated proteins
4. Autovalidate at 5% local FDR
5. Semi-tryptic search (nonspecific C-term) on validated proteins
6. Autovalidate at 5% local FDR
7. Semi-tryptic search (nonspecific N-term) on validated proteins
8. Autovalidate at 5% local FDR
9. Unknown modifications search (mass gap search) on validated proteins
10. Autovalidate at 5% local FDR

Each iteration has its own FDR analysis. The results from the iterations are combined based on the local FDR.

### Conclusions

- We present here a simple yet general method for calculating the local FDR
- We hope that this method will mitigate the practical (computational) disadvantages of the local FDR and lead to more widespread use of the local FDR in database search engines