Robust Method for Calculating the Local FDR for Database Search Results

ASMS 2010

Wilfred H. Tang, Agilent Technologies,
Santa Clara, CA

Robust Method for Calculating the Local FDR for Database Search Results

Introduction

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.

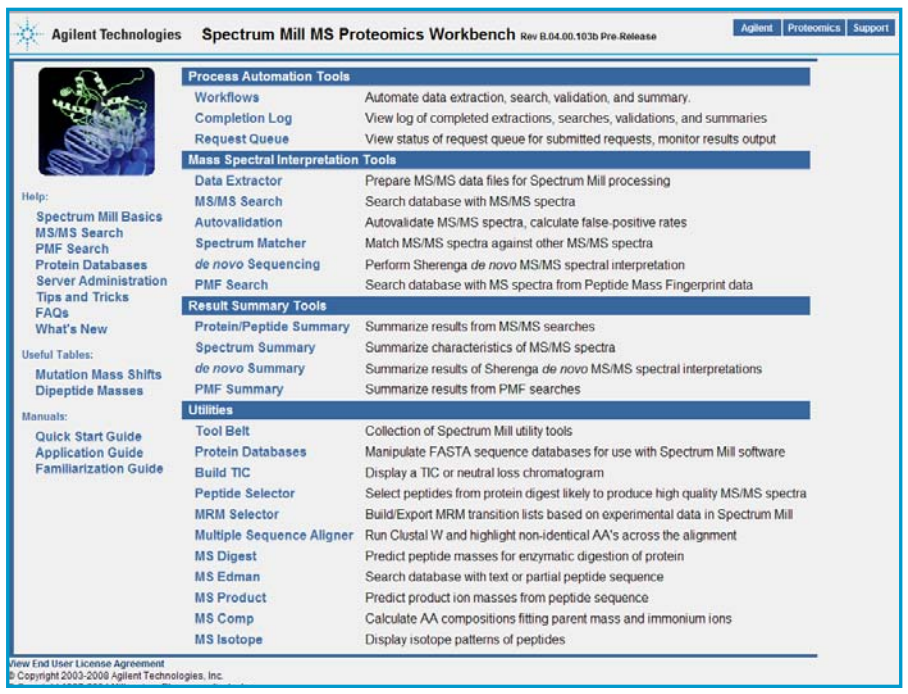
Advantages of Local FDR

- 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



Methods

Sample Preparation and LC/MS Analysis

Protein samples were reduced with DTT, alkylated with iodoacetamide, and digested with trypsin. The trypsinized HeLa cell lysate was fractionated into 24 fractions over the pH range 3 to 10 using an Agilent OFFGEL Fractionator.

Nanoflow LC/MS/MS was performed on an Agilent 6520 or 6530 Accurate-Mass Q-TOF with an HPLC-Chip/MS interface. Separations were done on an HPLC-Chip with a 75 µm x 150 mm analytical column and a 160 nL enrichment column.

Database Search

We use a pre-release version of Spectrum Mill MS Proteomics Workbench B.04.00 for performing database search.

Reversed sequences are used as decoys to estimate the rate of occurrence of false positive in the search results. Only the internal portion of each sequence is reversed – for example, SAMPLER is reversed to SELPMAR.

The list of IDs resulting from database search are ordered from best (highest score) to worst (lowest score). As the list is traversed from best to worst, let

- N be the total number of IDs traversed
- F be the cumulative number of false positives

Global FDR Calculation

The global FDR for each N is calculated as F/N.

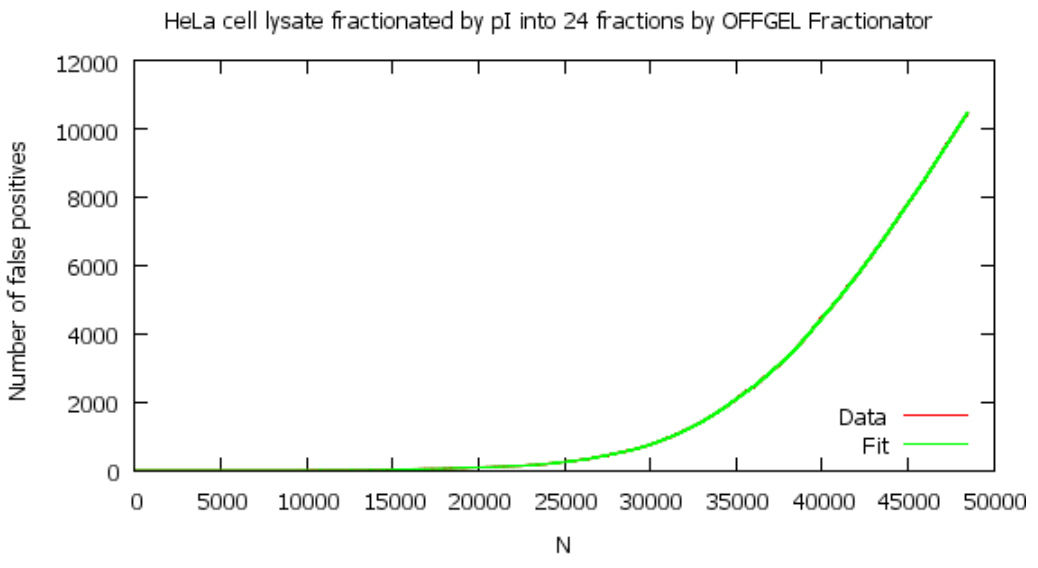
Local FDR Calculation

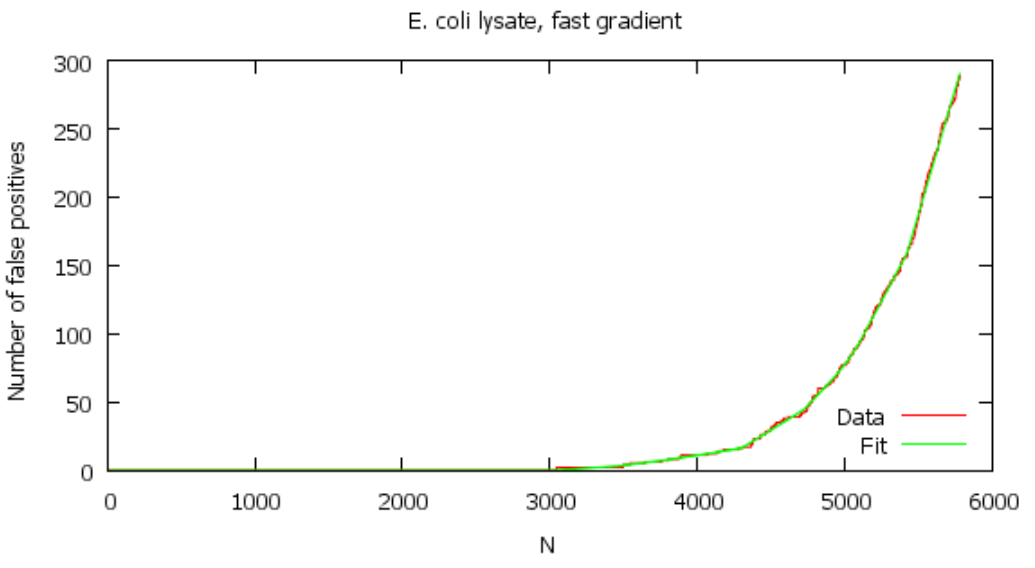
- In principle, the local FDR is calculated as the derivative dF/dN.
- In practice, the function $F(N)$ is "bumpy" → not smooth and requires smoothing and curve fitting in order to get a reasonable derivative.

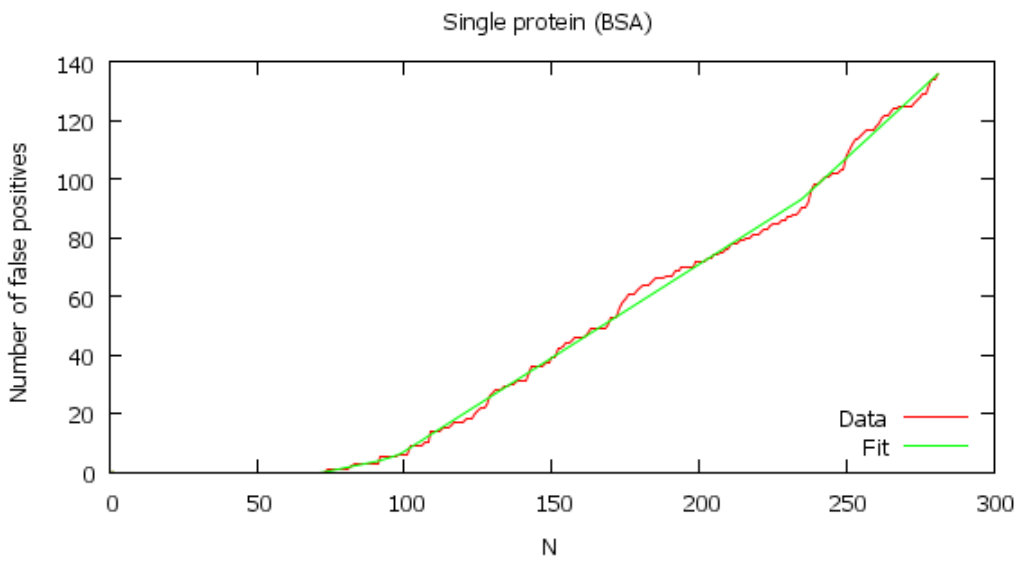
Smoothing out the bumps in F(N)

- Model $F(N)$ as a piecewise linear function
- Do a constrained least-squares fit. The constraint is an ordering constraint: the slope must be nondecreasing as N increases. Mathematically, this can be formulated as a convex optimization problem.
- We expect that this fitting method should work universally for calculating the local FDR for any database search engine. The only assumption (that the slope of $F(N)$ is nondecreasing as N increases) should hold true for any reasonable scoring function for peptide-spectrum matches.

Fitting F(N)



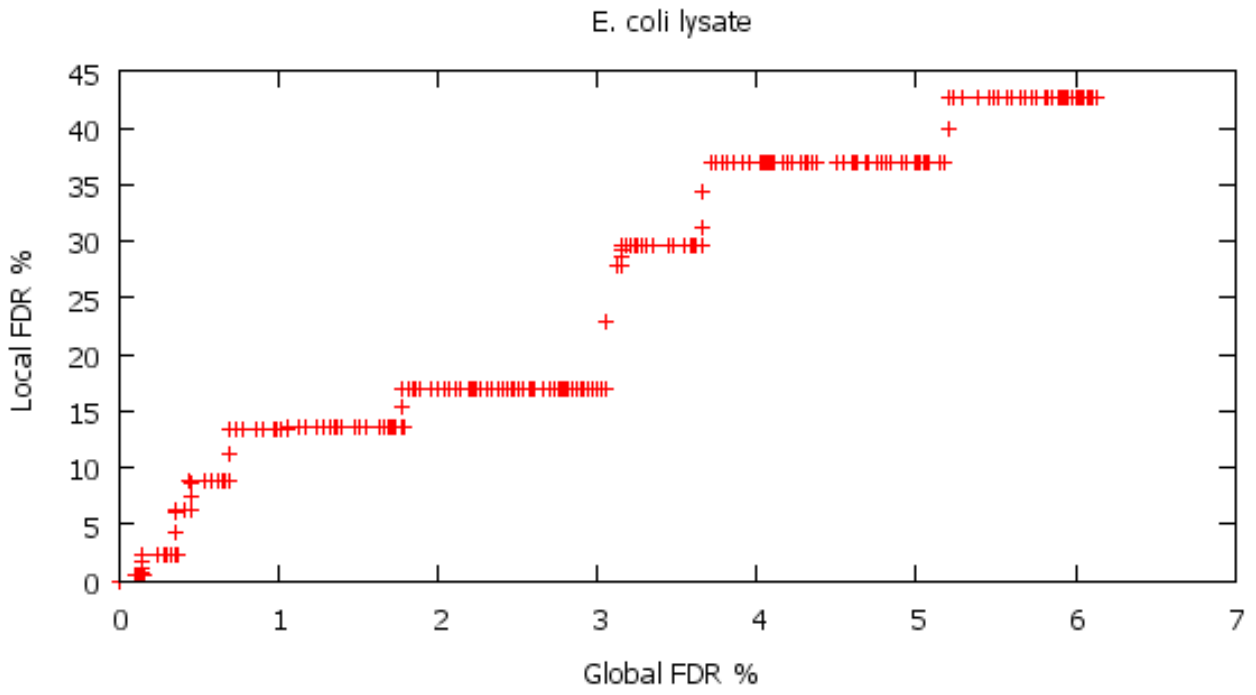




The key to successfully computing the local FDR is getting a good fit to *F*(*N*). We have tested our constrained least-squares fitting method on a variety of proteomics samples (of which 3 are shown here) and obtained decent fits in each case.

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).



Spectrum Mill FDR Analysis

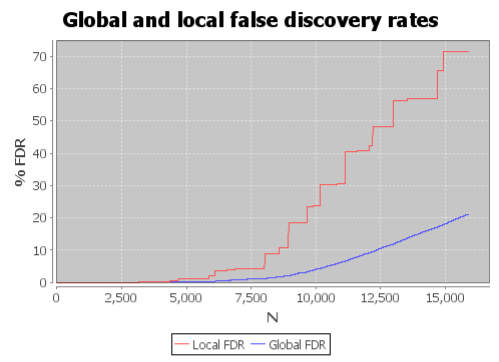
Example of Spectrum Mill's FDR analysis. Spectrum Mill reports both the global and local FDRs in tables and graphs and also shows the fit for *F*(*N*). The analysis is performed at 2 different levels – the spectral level and the distinct peptide level.

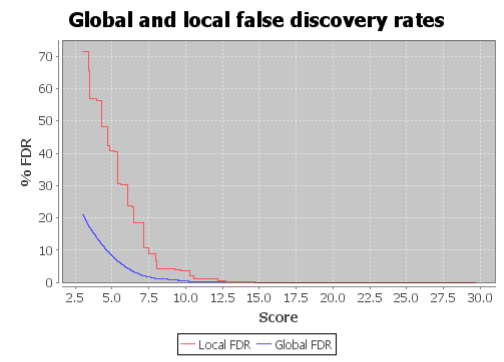
False Discovery Rate Search #1

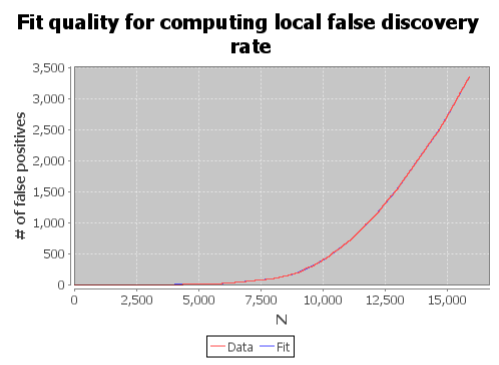
Spectral Level Analysis

	Number of spectra detected	
Critical FDR Value	Local FDR	Global FDR
1%	4687	7328
5%	8016	10427

	Score	
Critical FDR Value	Local FDR	Global FDR
1%	12.2012	8.77618
5%	8.02792	5.88024



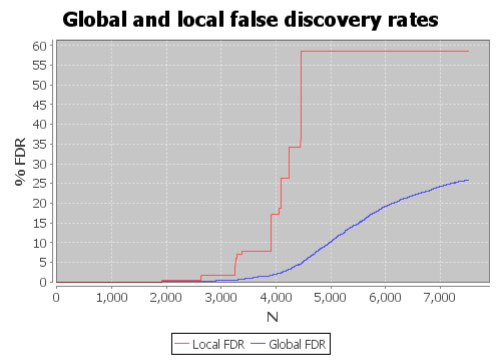


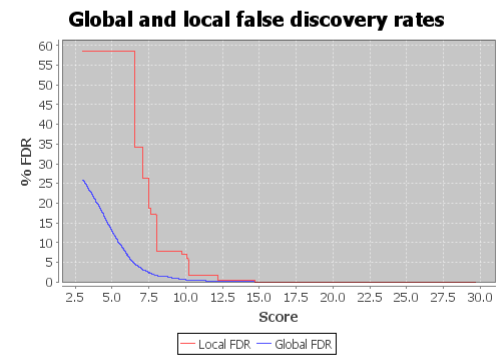


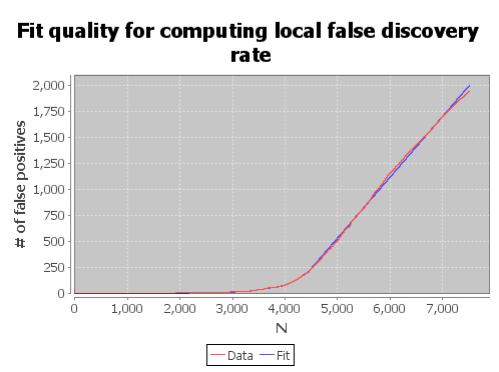
Distinct Peptide Level Analysis

	Number of distinct peptides detected	
Critical FDR Value	Local FDR	Global FDR
1%	2634	3526
5%	3258	4489

	Score	
Critical FDR Value	Local FDR	Global FDR
1%	12.2054	9.22742
5%	10.1862	6.43738







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

Agilent Spectrum Mill - Protein/Peptide Summary - (none)

Spectrum Mill | Summary Settings | Autovalidation | MRM Selector | MS/MS Search | Spectrum Summary | Build TIC | Workflows | Tool Belt | Help

Totals: 11681 5383

Group (#)	Subgroup (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Database Accession #	Protein Name
1	1.1	264	69	1213.00	85	P06733	Alpha enolase (EC 4.2.1.11) (2-phospho-D-glycerate

#	Filename	z	Score	Discriminant	Local FDR (%)	Global FDR (%)	FDR Search #	SPI (%)	Variable Sites	Sequence
1	HeLa-OGE-2GHz-fr12.8466.8709.2	2	26.43	26.4279	<0.1%	<0.1%	1	98.7		(K)VNQIGSVTESLQACK(L)
2	HeLa-OGE-2GHz-fr05.11373.11700.2	2	26.19	26.1853	<0.1%	<0.1%	1	99.0		(R)AAVPSGASTGIYEALELR(D)
3	HeLa-OGE-2GHz-fr17.10454.10811.3	3	26.18	26.1778	<0.1%	<0.1%	2	93.9	N362n	(K)LAQAnGWGVMVSHR(S)
4	HeLa-OGE-2GHz-fr13.8769.8966.2	2	25.16	25.1635	<0.1%	<0.1%	1	98.6		(K)VNQIGSVTESLQACK(L)
5	HeLa-OGE-2GHz-fr12.8475.8534.2	2	25.14	25.1405	<0.1%	<0.1%	1	98.2		(K)VNQIGSVTESLQACK(L)
6	HeLa-OGE-2GHz-fr17.7148.7222.2	2	25.04	25.0426	<0.1%	<0.1%	1	98.2		(R)IGAEVYHNLK(N)
7	HeLa-OGE-2GHz-fr02.16917.16965.3	3	24.91	24.9144	<0.1%	<0.1%	1	92.1		(K)DYPVVSIEDPFDQDDWGAWQK(F)
8	HeLa-OGE-2GHz-fr04.10659.10981.2	2	24.70	24.7034	<0.1%	<0.1%	1	95.9		(R)YISPDQLADLYK(S)
9	HeLa-OGE-2GHz-fr04.10785.10976.2	2	24.11	24.1069	<0.1%	<0.1%	1	98.2		(K)FTASAGIQVVGDDLTVTNPK(R)
10	HeLa-OGE-2GHz-fr04.9374.9442.2	2	23.84	23.8397	<0.1%	<0.1%	4	98.0		(F)TASAGIQVVGDDLTVTNPK(R)
11	HeLa-OGE-2GHz-fr11.8345.8604.2	2	23.84	23.8363	<0.1%	<0.1%	1	97.0		(K)VNQIGSVTESLQACK(L)
12	HeLa-OGE-2GHz-fr16.6171.6513.3	3	23.84	23.8389	<0.1%	<0.1%	1	89.4		(R)IGAEVYHNLK(N)
13	HeLa-OGE-2GHz-fr16.6198.6448.2	2	23.83	23.8267	<0.1%	<0.1%	1	94.9		(R)IGAEVYHNLK(N)
14	HeLa-OGE-2GHz-fr04.10200.10268.4	4	23.78	23.7846	<0.1%	<0.1%	2	80.0	N51n	(R)AAVPSGASTGIYEALELRDnDKTR(Y)
15	HeLa-OGE-2GHz-fr03.14432.14760.3	3	23.72	23.7222	<0.1%	<0.1%	1	100.0		(K)SFIKDYPVVSIEDPFDQDDWGAWQK(F)
16	HeLa-OGE-2GHz-fr03.12766.13064.3	3	23.69	23.6899	<0.1%	<0.1%	1	98.0		(K)DATNVGDEGGFAPNILENKEGLELLK(T)

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