ProRata: a Software Package with Improved Point and Interval Estimation of Protein Abundance Ratio for Quantitative Shotgun Proteomics

Nagiza F. Samatova[1,*], Chongle Pan[1,2], Guruprasad Kora[1], David L. Tabb[3,+], W. Hayes McDonald[2], Dale Pelletier[3], Greg Hurst[2] and Robert Hettich[2#]

[1]Computer Sciences and Mathematics Division, [2]Chemical Sciences Division and [3]Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN

Overview

Research Objectives:

Development of a free computer program for quantitative proteomics data analysis.

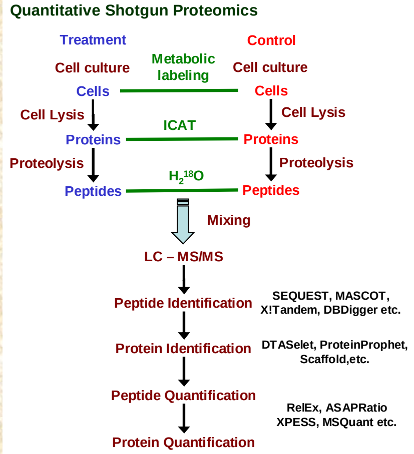
Key Features:

- Applicability to all existing isotopically labeling techniques.
- Robust peptide abundance ratio estimation with bias and variability scoring.
- Accurate protein abundance ratio estimation with confidence interval evaluation.
- Graphical user interface for interactive data interrogation and result validation.

ProRata
www.MSProRata.org

Introduction

Quantitative Shotgun Proteomics



Isotopic Labeling Methods:

- Metabolic labeling (^{15}N, ^{13}C and SILAC):
  - Mixing of the two samples at the earliest stage.
  - Requirement of cell growth with defined media
- Isotope-coded affinity tags (ICAT):
  - Mixing of the two samples before proteolysis.
  - Selected quantification of peptides with cysteine residues.
- $H_2^{18}O$ proteolysis with trypsin
  - Applicability to most proteome samples.
  - No correction to the bias to protein abundance ratios introduced during cell lysis and proteolysis

LC – MS/MS Measurement:

- Liquid chromatography separation: resolve peptides into chromatographic peaks.
- MS/MS scans: fragments peptides for amino acid sequence inference
- MS scans: measures the relative abundance of the isotopologue pairs of peptides

Data Analysis Challenges For Quantitative Proteomics:

- Challenge for peptide abundance ratio estimation: highly variable signal-to-noise ratios of the ion chromatograms requires a robust estimation algorithm and a confidence scoring method for expected estimation error.
- Challenge for protein abundance ratio estimation: Multiple peptides with inconsistent log-ratios and different estimation confidence from a protein requires a rigorous algorithm for "weighing" the peptides to yield an accurate point estimate and a precise confidence interval estimate.

Peptide Quantification

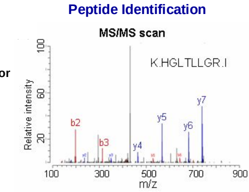
Estimation of Peptide Abundance Ratio and Profile Signal-to-Noise Ratio

The data analysis pipeline for quantitative proteomics

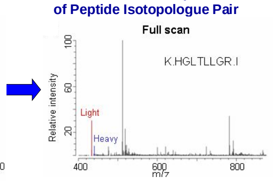
Novel algorithms employed in ProRata:

- Parallel paired covariance for chromatographic peak detection
- Principal component analysis for peptide abundance ratio estimation and scoring

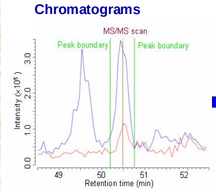
Peptide Identification



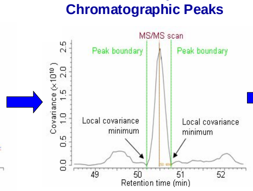
Detection of Mass Spectral Peaks of Peptide Isotopologue Pair



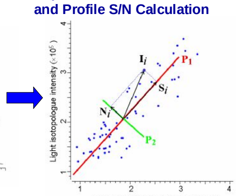
Extraction of Selected Ion Chromatograms



Detection of Peptide Chromatographic Peaks



Peptide Abundance Ratio Estimation and Profile S/N Calculation



Parallel Paired Covariance

Principal Component Analysis

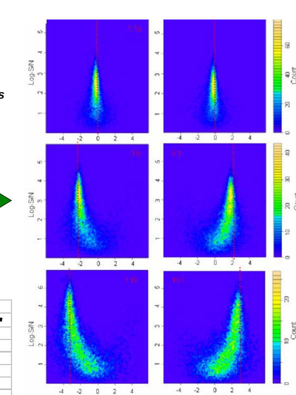
The algorithms were tested with six standard mixture of isotopically labeled *Rhodopseudomonas palustris* proteomes.

Experimental Material and Methods:

1. Preparation of two identically grown metabolically labeled *R. palustris* cultures, one with ^{14}N labeling and the other one with ^{15}N labeling.
2. Mixing of the two extracted proteomes at following ratios of total protein amount: ^{14}N : ^{15}N = 10:1, 5:1, 1:1, 1:5 and 1:10.
3. Analysis of the standard mixtures with 12-step MudPIT on a LTQ-MS instruments.

Heatmap histogram of log-ratio and log-S/N:

Profile S/N is correlated with the variability and bias of peptide log-ratio estimation.



Summary of Peptide Quantification Results for the Standard Mixtures

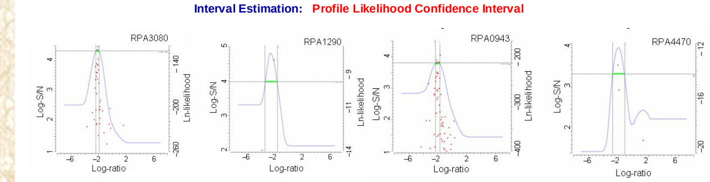
Standard Mixtures		Peptide Counts				Log-ratio		Log-S/N	
^{14}N:^{15}N	Log-ratio	^{14}N ID	^{15}N ID	SIC#	Quantified	Median	AAD*	Median	AAD*
1:1a	0.0	13,766	11,665	17,574	11,818	-0.17	0.67	2.17	0.58
1:1b	0.0	13,975	12,230	18,472	12,958	-0.16	0.69	2.15	0.56
5:1	2.3	23,527	5,122	23,256	14,583	1.66	1.04	2.29	0.73
1:5	-2.3	5,676	18,037	18,855	12,453	-1.99	0.94	2.44	0.78
10:1	3.3	24,257	2,725	22,770	12,914	2.20	1.37	2.31	0.85
1:10	-3.3	3,167	21,396	20,945	12,910	-2.80	1.39	2.46	0.91
Average		--	—	20,312	12,956	—	1.02	2.30	0.74

Protein Quantification

Estimation of Protein Abundance Ratio from Peptide Abundance Ratio Scored with Profile S/N

Profile Likelihood Algorithm: Point Estimation: Maximum Likelihood Estimation

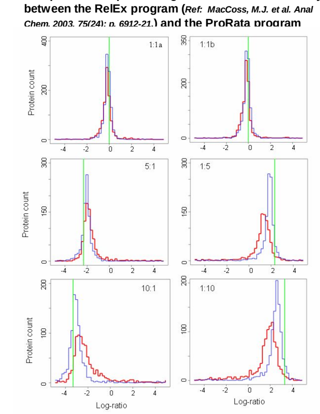
Interval Estimation: Profile Likelihood Confidence Interval



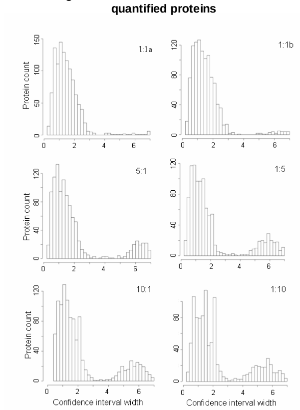
Algorithm features: Weighted peptide log-ratios | Correction of peptide log-ratio bias | Suppression of outliers

Benchmark of Profile Likelihood Algorithm Performance with Standard Mixtures

Comparisons of protein log-ratio estimation accuracy between the RelEx program (*Ref: MacCoss, M.J. et al. Anal Chem. 2003. 75(24): p. 6912-21.*) and the ProRata program



Histogram of confidence interval width for the quantified proteins



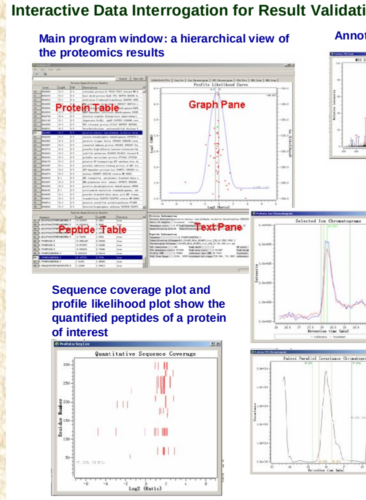
Summary of Protein Quantification Results for the Standard Mixtures

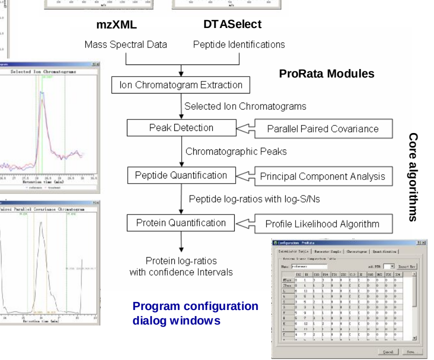
Hypothesis testing with confidence interval estimates

Standard mixture		Protein count		Log-ratio point estimation		Confidence interval estimation		Hypothesis testing	
^{14}N:^{15}N	Log-ratio	Identified	Quantified	Median	AAD*	Median width	Confidence level	Significance	Power
1:1a	0.0	1,392	1,117	−0.2	0.318	1.4	93%	8%	---
1:1b	0.0	1,348	1,071	−0.2	0.361	1.4	92%	10%	---
5:1	2.3	1,384	1,054	1.8	0.481	1.4	90%	---	94%
1:5	-2.3	1,263	1,024	−2.1	0.390	1.4	92%	---	97%
10:1	3.3	1,475	1,096	2.5	0.561	1.6	88%	---	96%
1:10	-3.3	1,312	1,000	−3.1	0.639	1.6	87%	---	98%
Average		1,362	1,060	---	0.458	1.5	90%	9%	96%

Graphical User Interface

Interactive Data Interrogation for Result Validation





Conclusions

- Parallel paired covariance algorithm is developed for robust peak detection in low-S/N selected ion chromatograms from peptide isotopologue pairs
- Principal component analysis algorithm allows accurate peptide abundance ratio estimation and the estimation variability and bias are correlated with profile signal-to-noise ratio.
- Profile likelihood algorithm gives maximum likelihood point estimation and profile likelihood confidence interval estimation for protein abundance ratios.
- Graphical user interface of ProRata enables manual validation of the protein quantification results
- The ProRata program is freely available for download at http://www.MSProRata.org

Acknowledgements

- Research at ORNL sponsored by the U.S. Department of Energy, Office of Biological and Environmental Research, Genomics:Genomes to Life, under contract No. DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed and operated by UT-Battelle, LLC.
- C.P. would like to acknowledge the financial support from Genome Science & Technology Graduate School.

- Contact: * N.F. Samatova samatovan@ornl.gov and # R.L. Hettich hettichrl@ornl.gov
- + Current address of D.L. Tabb: Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232