An integrative scoring scheme for protein identification in tandem mass spectrometry experiments

Smriti R. Ramakrishnan, Christine Vogel, John T. Prince, Zhihua Li,
Edward M. Marcotte, Daniel P. Miranker

Motivation:

An understanding of the spatial and temporal changes in mRNA and protein expression in organisms is a fundamental problem in biology. While mRNA expression levels are routinely measured on large scale, existing methods of protein identification in complex samples (e.g. Western blotting, 2D-gel electrophoresis, GFP-tagging) are very labor-, time- and resource-intensive. Mass-spectrometry (MS) based MudPIT proteomics provides a fast, high-throughput alternative.

However, standard instruments often identify only few hundred proteins in a complex protein sample. Moreover, in an MS/MS experiment, identification of a protein can be hindered by low abundance, post-translational modifications or chemical properties that interfere with efficient ionization of the protein's sequence. As a result, the protein is associated with a raw MS/MS identification score below a given confidence threshold, e.g. 5% False Discovery Rate (FDR), marking it as absent. Here, we boost identification of proteins by integrating information on mRNA expression levels with MS/MS protein identification scores in a Bayesian framework, and we correctly identify proteins as being present. We establish and validate our system called MS-BOOST in yeast, reporting several hundred additional proteins at high confidence.

Methods:

We formulate a new integrative Bayesian identification probability score, the MS-BOOST score, by combining priors learnt from both *inferential* and *direct* evidence of protein presence. In our case, *direct* evidence of protein presence originates from a raw MS/MS protein identification score [1] and *inferential* evidence is derived from associated mRNA expression data.

We formulate a Bayesian score $P(k|S_k, M_k)$, the probability that protein P_k exists in the sample given its raw protein identification score in an MS/MS experiment (S_k) and its associated mRNA abundance (M_k). We estimate the conditional probabilities $P(k|S_k)$ and $P(k|M_k)$ required to generate the integrated score using MS/MS protein identification scores from a) published LCQ MS/MS analysis on a yeast sample grown in rich medium [2], b) associated mRNA abundances [2] and c) non-MS-based protein identification datasets ([3], [4],[5]).

Results:

MS-BOOST nearly doubles the number of proteins identified in the yeast sample, i.e. among proteins that have associated mRNA abundances we identify 741 instead of the original 396 proteins at 5% FDR. We are able to validate 92% of the new predictions with eight independent datasets which are based on MS ([6],[7],[8],[9]) and other methods ([3],[4],[5],[10]). The expression of the remaining 8% (25) newly identified proteins is biologically meaningful in all cases except for one. Our MS-BOOSTed protein identification scores outperform raw MS/MS identification scores in cross-validation

based recall-precision curves using a ground truth set of expected proteins ([3],[4],[5]). For example, MS-BOOST scores result in an F-measure (harmonic mean of precision and recall) of 86.2% 5% FDR probability cutoff versus 60.9% when using raw probabilities.

In a second application of the generic score formulation and learnt probability priors of MS-BOOST to an LC/LC-MS/MS analysis of yeast on OrbiTrap [9], we identify ~500 additional proteins at 5% FDR, observing a total of >2,500 proteins and validating ~75% of the new proteins using the validation sets mentioned above. We are currently testing MS-BOOST on *E.coli* proteomics data using the trained eukaryotic model, and we are compiling experimental datasets to build and validate the prokaryotic model.

Conclusions:

We present a novel method, MS-BOOST, which integrates prior evidence from mRNA coding data with raw MS/MS identification scores to boost protein identification in MS/MS shotgun analysis, and show that several hundreds of additional proteins can be reliably identified in data from a single experiment. We use MS-BOOST to examine the yeast and *E.coli* proteome, e.g. we aim to determine the upper bound of number of proteins expressed in steady state. Our probabilistic framework is generic and can easily be applied to other organisms and MS/MS scoring schemes.

References

[1] Nesvizhskii, A.I. et al, A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry, *Anal. Chem*. 2003, 75, 4646-58.

[2] Peng Lu, Christine Vogel, Rong Wang, Xin Yao and Edward M. Marcotte, Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation, *Nature Biotechnology* 2006, **25**, 117–24.

[3] Newman, J.R. et al. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* (2006).

[4] Ghaemmaghami, S. et al. Global analysis of protein expression in yeast. Nature 2003, **425**, 737-41

[5] Futcher, B. et al, A Sampling of the Yeast Proteome, Mol Cell Bio 1999, 19(11):7357-68.

[6] Lyris MF de Godoy, Jesper V Olsen, Gustavo A de Souza, Guoqing Li, Peter Mortensen and Matthias Mann, Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system, Genome Biology, 2006, 7:R50.

[7] Peng, J. et al, Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome, J. Proteome Res. 2006; 5(9); 2372-79

[8] Washburn, M.P. et al, Large-scale analysis of the yeast proteome by multidimensional protein identification technology, Nature Biotechnology 2001, 19(3):242-7.

[9] Open Proteomics Database, http://bioinformatics.icmb.utexas.edu/OPD

[10] Chi, A. et al, Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry, PNAS 2007, 104(7):2193-8.