

Corra, APLM and TIQAM: computational software tools for discovery and validation of biomarker candidates

Mi-Youn Brusniak¹, Kelly Cooke¹, Alexander Schmidt², Simon Letarte¹, Carey Sheu¹, Julian Watts¹, and Ruedi Aebersold^{1,2}.

¹Institute for Systems Biology, 1441 North 34th Street, Seattle WA 98103, USA, ²Institute of Molecular Systems Biology, ETH Zurich HPT E 78 CH-8093 Zurich, Switzerland.

Overview

- Corra is software for discovering biomarkers by detecting differentially expressed features/peptides by label-free LC-MS quantification proteomics.
- TIQAM (Targeted Identification for Quantitative Analysis by MRM) is a suite of software which was developed for supporting hypothesis driven biomarker validation by SRM (Selected Reaction Monitoring).
- The candidate peptide or protein lists for TIQAM can be generated directly from Corra.
- Corra :TIQAM together created a tool set for supporting a workflow of discovery and followed by verification of biomarker candidates.
- To illustrate the workflow for Corra:TIQAM, we show the application of the tools for human disease biomarker discovery and verification.

Introduction

We introduce a complete software suite named Corra:TIQAM as open source software tools for discovery and followed by verification of biomarker candidates and other proteomic applications in which multiple samples are being quantitatively compared. The Corra software component is for generating target candidate lists via label-free LC-MS quantification proteomics. The TIQAM software component is for assisting the selection of candidate transitions and their validation for Selected Reaction Monitoring (SRM) proteomic experiments in large sample series. Here, we introduce and illustrate the complete Corra:TIQAM software suite applied to the discovery and verification of biomarker candidates, and include a brief discussion of planned future work.

Methods

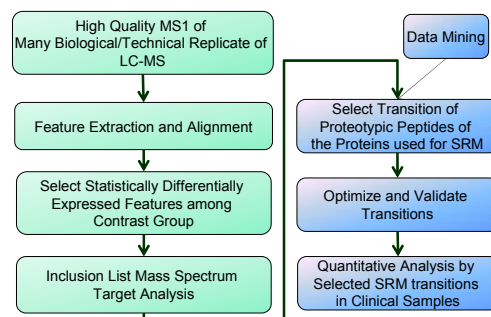
Corra¹

Corra is a single, user-friendly, informatic framework, that is simple to use and fully customizable, for the enabling of label free LC-MS-based quantitative proteomic workflows of any size, able to guide the user seamlessly from MS data generation, through data processing, visualization, and statistical analysis steps, to the identification of differentially abundant or expressed candidate features for prioritized targeted identification by subsequent MS/MS. A goal of Corra was to enable the integration of multiple and disparate LC-MS data analysis tools, and integrate them, seamlessly, with common statistical packages to allow for better comparison between differently-processed datasets, via the addition of statistical measures of confidence and error rates. The integration of tools was achieved via AMPL (Annotated Putative peptide Markup Language) and the various parsers, and in the current build of Corra, we have implemented SpecArray², SuperHirn³, msBID⁴. Corra v1.5 is an open source software currently distributed through SPC (Seattle Proteome Center) website <http://tools.proteomecenter.org/Corra/corra.html>.

TIQAM⁵

TIQAM is developed as a suite of software tools to support targeted identification and quantification using Selected Reaction Monitoring (SRM) mass spectrometry technology. More specifically, TIQAM software provided user friendly interfaces to assist process of peptide selection, transition generation and validation. TIQAM is also an open source software and currently distributed through SPC (Seattle Proteome Center) website <http://tools.proteomecenter.org/TIQAM/TIQAM.html>

Corra:TIQAM Workflow



Results

Getting Candidate Protein List using Corra

Eleven matched tumor-control Human tissues were obtained and N-Glycopeptide enrichment⁶ was performed on all samples. The 24 samples were run, in triplicate, on an LTQ-FT collecting MS1 and MS2. Due to heterogeneity of glycoproteins in each patient samples, we randomized LC-MS runs within a given patient and using SuperHirn/Corra was used to separately align the 6 FT-MS runs from each matched tumor-control tissue pair, for all 11 patients. The differentially abundant peptide list was thus generated on a per patient basis. A cut-off for differential abundance confidence was set at log Odds ≥ 2.2 ($\approx 90\%$ chance of observed difference being non-random). The lists were used as an inclusion list for targeted MS runs. Data from the DDA and targeted sequencing were combined, data searched and peptide sequence IDs mapped back onto aligned peptide features, and aligned peptide IDs mapped onto IPIs and Gene Symbols. Some of the candidate peptide/proteins are used for transition optimization process using TIQAM for further SRM verification.

Number of Patients	Gene	Protein	Peptide
10	CECR1	cat eye syndrome chromosome region, candidate 1	VQDVTEFDDSLR
10	FAP	fibroblast activation protein, alpha	SVDSANVGLSPDR
10	FKBP10	FKBP5 binding protein 10, 65 kDa	YHYDYLTLGDTSFDTSYSK
10	LGALS3BP	MAC-2 binding protein precursor	ALGFEDATQALGR
11	POSTN	Periostin, Osteoblast-specific factor 2	EVDDTLVNELK
10	THBS1	thrombospondin 1	VVDSTTGPGEHLR

Table 1. Some of Example Protein Lists used in Optimizing Transitions for further verification.

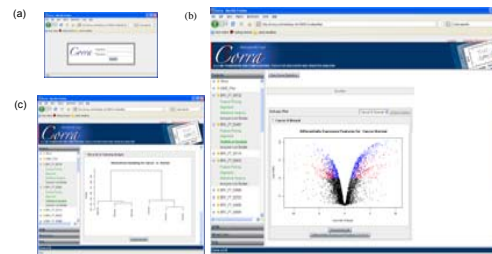


Figure 1. (a) Corra was developed as an intranet-based web application to handle several users, and to help users organize their processes by project. Input files are in mzXML format. Color of projects and pipeline steps indicates the preprocessing status. Yellow indicates running process, green indicates completed process, red indicates failed process. (b) Volcano plot shows the feature that are differentially expressed, blue dots indicate features which were aligned across all samples, and red dots indicate features with some missing values among aligned samples.

Selection & Optimization of Transitions using TIQAM

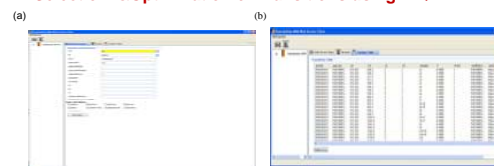


Figure 2. TIQAM-PeptideAtlasClient module helps users select and filter potential transitions by looking up measured proteotypic peptides by interfacing PeptideAtlas⁷ consensus spectra libraries. (a) shows the query selection interface, and (b) shows the results from the search in a table format for user to sort and filter the transitions.

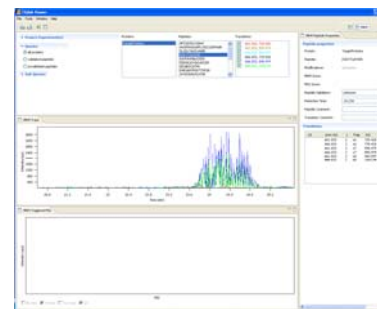


Figure 3. TIQAM-Viewer module takes mzXML, pepXML and transition files, and visualizes SRM traces and spectra. It provides a graphical environment to analyze SRM triggered MS2 experiments, and to let users annotate the transitions as valid or not valid, in order to design the next iterative process of the SRM measurements.

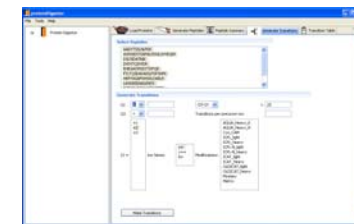


Figure 4. TIQAM-Digester module helps users generate *in silico* transitions from theoretical digestion of the proteins in a given input FASTA file. It also allows users to import properties of the transitions in order to annotate them, and to help the prioritization of the transitions to be measured.

Discussion and Conclusions

We presented Corra:TIQAM as a proteomics workflow and software tool set that can utilize discovery and verification of candidate proteins for biomarker or general proteomic study. Corra's open architecture allows the computational biologist community to freely adapt their preferred modules to Corra's pipeline. The Corra framework and computational tools for MS1 data generates differentially abundant MS1 feature lists for targeted MS2 analysis. The follow-up targeted MS2 analyses enable the identification of differentially abundant features, thus generating candidate lists for biomarker discovery. Corra has already been used in various biological studies.^{1,8} TIQAM software is designed to assist our SRM experimental workflow, applied for *Streptococcus pyogenes* virulence factor study². Current TIQAM tool sets are valuable for manual optimization of transitions but it's operator intensive data validation. Therefore, we have developed other SRM workflows to support more scalable and high throughput SRM studies in the near future.

References

1. Brusniak et al. (2008) Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinformatics*, 9, 542
2. Li et al. (2005) A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol. Cell. Proteomics*, 4, 1328-40.
3. Muller et al. (2008) An Assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J. Proteome Res.*, 7, 51-61
4. Huang et al. (2008) MS-BID: a Java package for label-free LC-MS based comparative proteomic analysis. *Bioinformatics* accepted
5. Lange et al. (2008) Targeted quantitative analysis of *Streptococcus pyogenes* virulence factors by multiple reaction monitoring. *Mol. Cell. Proteomics*, 7, 1489-1500.
6. Zhang et al. (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nature Biotechnology*, 21, 660-666
7. Deutsch et al. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *Genome Biol.*, 9, R6.
8. Letarte et al. (2008) Differential Plasma Glycoproteome of p19ARF Skin Cancer Mouse Model Using the Corra Label-Free LC-MS Proteomics Platform. *Clinical Proteomics*, 4, 105-116.

Acknowledgements

This work was supported in part with federal funds from the National Heart, Lung, and Blood Institute Seattle Proteome Center (contract No. N01-HV-8179 to R.A.), the National Cancer Institute (contract No. N01-CC-12400 to J.W.), and National Institute of Diabetes & Digestive & Kidney Disease (grant No. 1R21DK71275 to J.W.).