

Corra: Computational Tools for Discovery and Targeted Mass Spectrometry Application to Candidate Biomarker Identification for Human Type 2 Diabetes

Mi-Youn Brusniak¹, Simon Letarte¹, Olga Vitek³, David Campbell¹, Lukas Muller², James Eddes¹, 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 and ³Statistics and Computer Science Department, Purdue University, 150 N. University Street, West Lafayette, IN 47907

We introduce **Corra**: An LC-MS based framework and tools for both discovery and targeted MS approaches to candidate biomarker identification.

- LC-MS tool: LC-MS (MS1) feature identification in distributed batch mode and alignment of multiple runs.
- LC-MS common interface: XML schema called APML for capturing processed LC-MS data and interoperable interface among LC-MS tools as well as CorraStatistics.R module.
- CorraStatistics.R: normalization of aligned features and determination of differentially abundant features with B-statistic values for targeted analysis and supervised clustering.
- Target Analysis Tool: annotation of inclusion lists with peptide and protein identifications obtained via targeted tandem mass spectrometry (MS2).
- Classification tool: identification of features that yield a predicted model for unknown samples.

Introduction

We present the *Corra* framework and tools that enable processing of large pools of sample data in high throughput. *Corra* consists of computational tools for both discovery and targeted MS approaches to candidate biomarker identification.

Corra tools include those for label free MS1 quantification, feature alignment, statistical analysis of differentially abundant features, annotation of MS1 features with MS2 data and the generation of inclusion lists for targeted MS. *Corra* also includes a common interface, APML (Annotated Putative peptide Markup Language). APML is a XML schema to capture processed MS1 data, including feature extraction, alignment and profiling data. APML facilitates interoperability of existing and new tools and processed data management.

Corra is currently used for posttranslational modification studies as well as biomarker discovery studies at ISB and IMSB. In this poster, we describe the *Corra* APML schema and application of its computational tools to generate an annotated candidate biomarker list for human type 2 diabetes.

Data sets

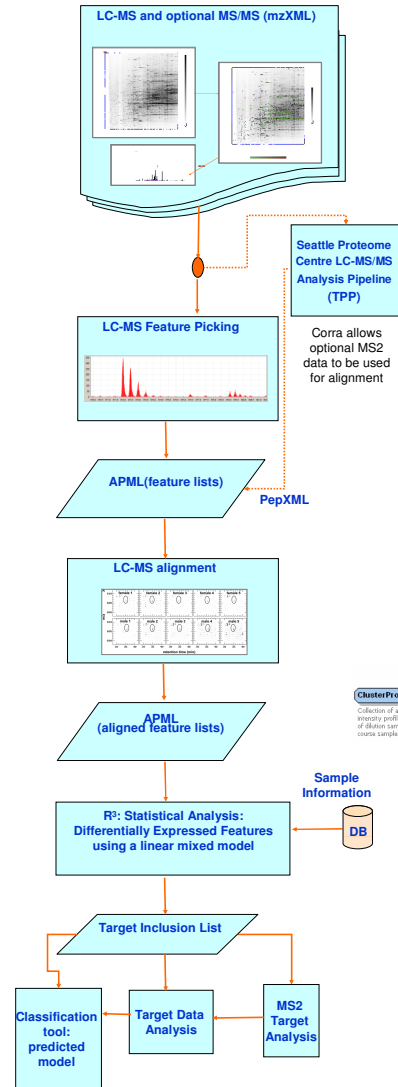
35 human plasma samples were obtained that were classified as being from individuals having either normal glucose tolerance (NGT) or impaired glucose tolerance (IGT), or being newly diagnosed with type 2 Diabetes (DB). The classifications were made via the oral glucose tolerance test, the diagnostic standard for diabetes. The NGT group consisted of 13 patients with blood glucose levels ranging from 54 to 98 mg/dl glucose 2h post-glucose challenge after fasting (NGT is ≤ 100 mg/dl). The IGT group also consisted of 13 patients, ranging from 142 to 191 mg/dl glucose (IGT is >100 and < 200 mg/dl). The DB group consisted of 9 patients ranging from 202 to 279 mg/dl glucose (DB is diagnosed at ≥ 200 mg/dl). Glycopeptide enrichment¹ was performed on all samples and MS1 profiles generated for each individual preparation, in triplicate, on a Bruker ESI-TOF spectrometer. Targeted MS2 analyses were on a LTQ-FT spectrometer using 4 each of the NGT and DB samples selected at random.

Corra Performance

High-Throughput Multi-threaded and Distributed Process

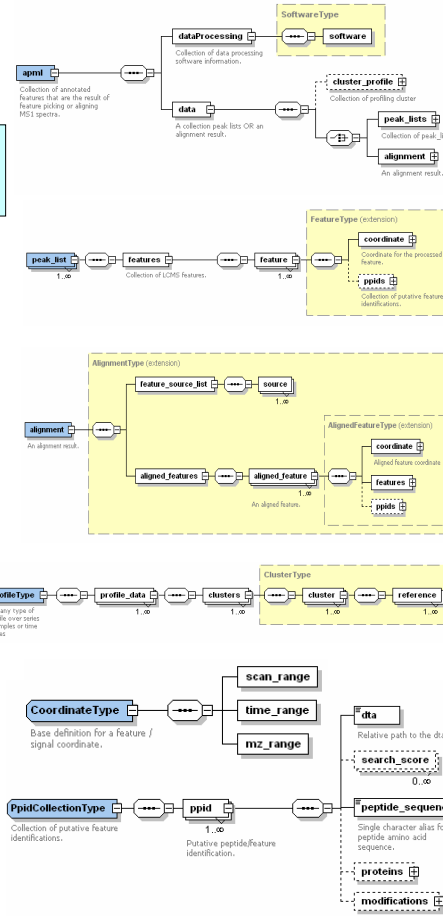
Corra utilizes an adaptation of an existing LC-MS tool (SpecArray²) for use in a multi-threaded and distributed computing environment. Feature extraction for the 105 MS1 runs (~3 GB mXML file per run) took ~25 hours using a six dual core, dual processor AMD Opteron 275, 2.2GHz, 1MB level 2 cache hardware configuration.

Methods: Corra workflow



Corra APML

Corra introduces a common interface called APML (Annotated Putative peptide Markup Language). APML is a XML schema to capture processed MS1 data, including feature extraction, alignment and profiling data. APML facilitates interoperability of existing and new tools and processed data management. The APML parser library is written Java 6 using SAX and StAX APIs.



Results

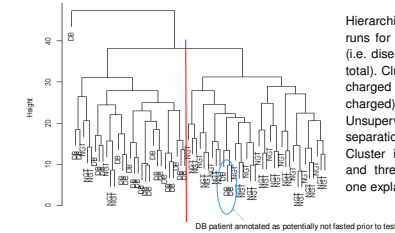


Figure 1

Hierarchical clustering of triplicate MS1 runs for the 13 NGT and 9 DB patients (i.e. disease vs. control: 66 LC-MS runs total). Clustering utilized the 588 multiply charged features (i.e. excluded singly charged) aligned across all 66 runs. Unsupervised clustering shows good separation between the two groups. Cluster indicates one bad LC-MS run and three miss-classified patients with one explainable case.

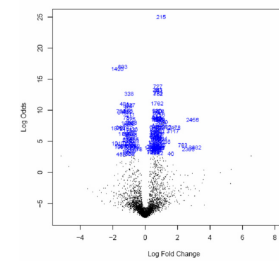


Figure 2

CorraStatistics.R generated "Volcano" plot using a linear mixed model (Limma⁴). The plot illustrates detection of differential abundance of the 4240 aligned features that aligned across at least three MS1 runs. x-axis: observed log fold change in feature intensity. y-axis: B-statistics log Odds of differential expression obtained for individual features. Dots associated with feature IDs in blue have higher than a log Odds value of 3.0. The target list was generated for the top 400 aligned features generated by CorraStatistics.R.

Table 1

Supervised clustering (k-TSP⁵) yielded a predictive model with a 91% LOOCV accuracy. The final 5 pairs of features were selected in the predictive model from the target list. The model classified the 13 IGT individuals thus: 3 as DB, 7 as NGT and 3 as undecided.

Index	m.z	rt	q	P.Val	B	protein	NCBI Gene Description
215	545.299	37.715	2	0.0000	25.03	IP100188529,IP100180424,IP100375327,IP100749459	chromosome 1 open reading frame 125
1420	522.789	33.42	2	0.0000	16.68	IP100232019	sex hormone-binding globulin
727	626.297	50.643	2	0.0000	13.79	IP100452803,IP100549376	activating signal integrator 1 complex subunit 2
343	1178.527	38.235	2	0.0000	5.09	IP100479708,IP100549291,IP100748158,IP100784931	immunoglobulin heavy constant mu
1731	764.815	40.694	2	0.0001	3.73	IP100783455	chromosome 11 open reading frame 41
3539	537.779	28.457	2	0.0024	0.70	IP100322434,IP100745872	albumin
856	597.797	44.55	2	0.0097	-1.66		
935	722.06	71.964	3	0.0001	4.15	IP100478003	alpha-2-macroglobulin
1083	490.571	16.211	3	0.0020	0.23	IP100228413	inter-alpha (globulin) inhibitor B3
212	487.242	30.393	3	0.0032	-0.34	IP100431645,IP100477597,IP100478493,IP100607707,IP100641737	haptoglobin-related protein

Conclusions

The *Corra* framework and computational tools for MS1 data can process large data sets and generate differentially abundant MS1 feature lists for targeted MS2 analysis. The APML interface facilitates interoperability of the underlying MS1 and bioinformatics tools. The follow-up targeted analyses enables the identification of differentially abundant features, thus generating candidate lists for biomarker discovery. *Corra* is a novel and functional computational framework enabling the implementation and interchanging of various software tool sets for MS data analysis in a high throughput proteomic workflow environment. *Corra* is currently in use at both ISB and IMSB for posttranslational modification studies as well as biomarker discovery studies for cancer and type 2 diabetes.

References

- 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
- 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.
- Geniflaman et al. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer New York.
- Smyth G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3.
- Tan et al. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 19, 84-90.