An improved data processing pipe-line for comprehensive H-NMR and X/MS -omics data

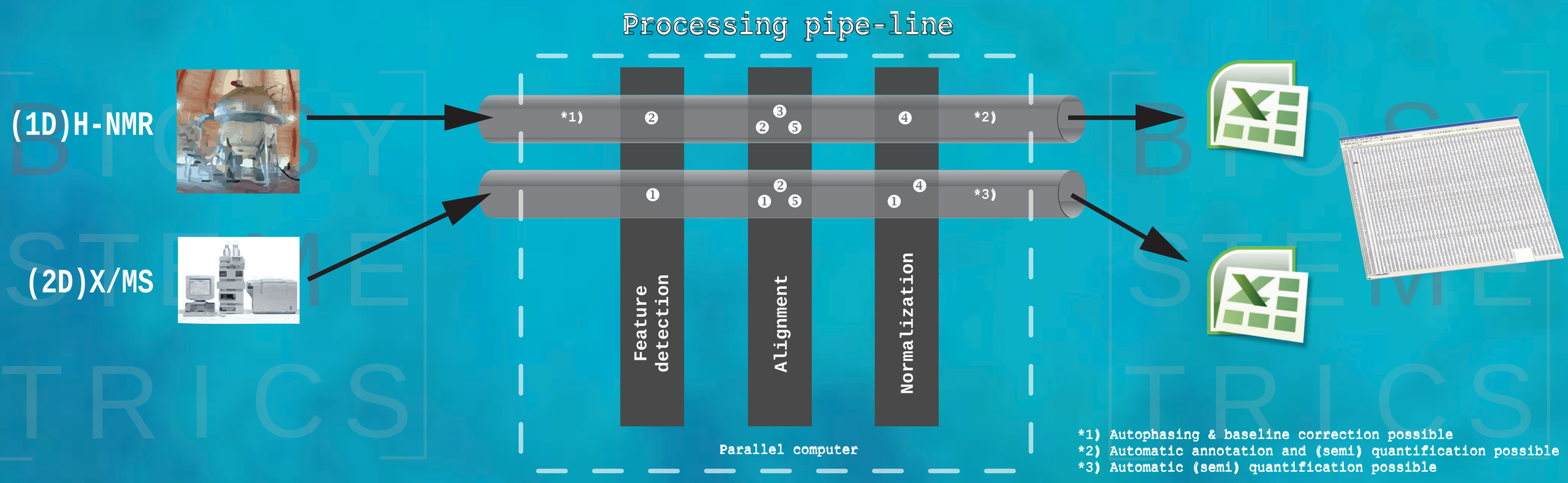
Ralf J.O. Torgrip, K. Magnus Åberg, Erik Alm

Stockholm University, Dept. of Analytical Chemistry, BioSysteMetrics Group, SE-106 91, Stockholm, Sweden.

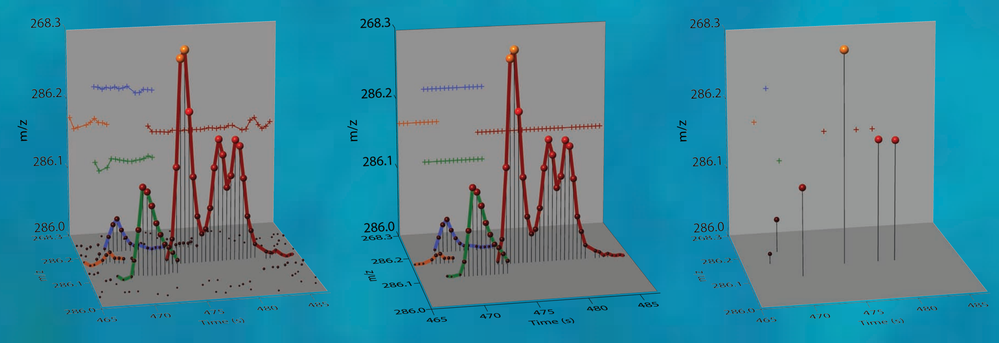
In the post acqusition analysis of comprehensive -omics data the pre-processing pipe-line is crucial to extract the maximum possible amount of information from the data.
Here we show a processing pipe-line for (1D)H-NMR and (2D)X/MS data comprising; feature detection, inter-sample feature alignment and internal-standard free normalization. I particular we show that employing new processing methods it is possible to:

a) significantly increase the number of detected features in the data,
b) significantly increase the confidence in the resulting data due to inter-sample feature alignment,
c) remove data- and processing-bias due to the low number of meta-parameters associated with the processing steps.

Combining these features we gain an increase of the extractable information content with at least a factor of four (>4!) measured as the number of corresponding peaks over a multi-sample data-set. We present a processing strategy based on matched filters, Kalman tracking, histogram normalization and Hough transform alignment that outperforms today's state-of-the-art processing schemes. Furthermore, the resulting data comes in tabular format making statistical tests such as ANOVA and t-tests straightforward to implement.

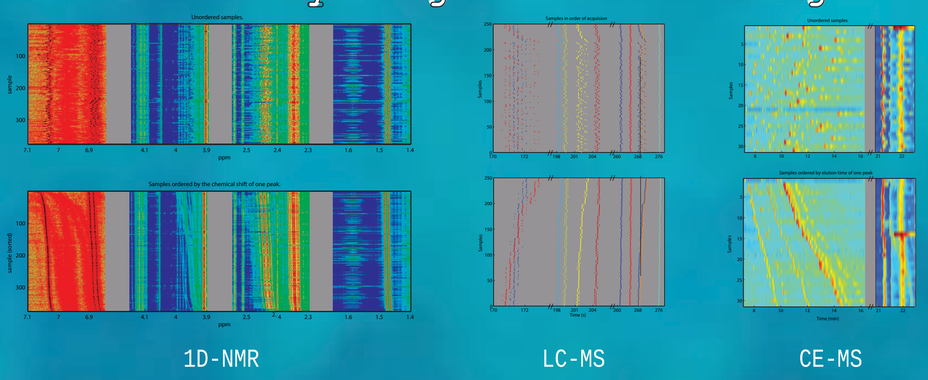


TracMass X/MS detection/alignment ❶



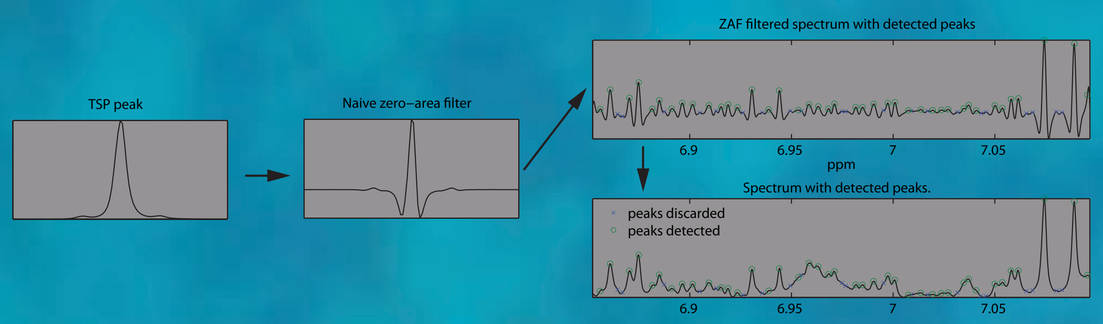
TracMass ❶ feature detection/alignment. Pure Ion chromatograms (PICs) are collated, aligned and peak-detected using Kalman Filter tracking - analogous to airplane and ICBM tracking in radar data. The method has only one user defined parameter - the minimum number of scans/peak. The result is that each true peak is extracted by its time, intensity (integral) and weighted m/z. ***No*** thresholds, binning or other cutoffs are used resulting in that the ***full*** latent information content of the sample is extracted.

Generalized Fuzzy Hough Transform alignment ❷ ❸ ❺



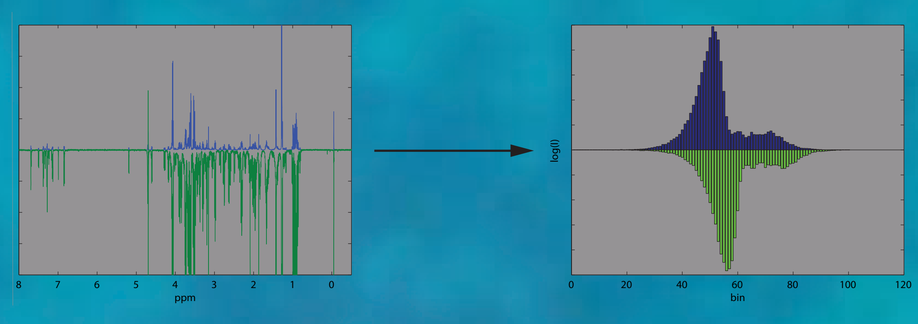
The Generalized Fuzzy Hough Transform alignment ❷❸❺ is the first alignment method that can align peaks that changes their respective order. Furthermore, the method generalizes over different measuring platforms such as NMR, LC/MS, CE/MS. The GFHT is derived from image analysis and is based on the fact that peak location is not a random feature but but in fact deterministic. The GFHT has one user defined parameter which is easily set - the number of latent shift patterns occurring in the current data set. This parameter is derived from the data.

Naive zero area-filter peak detection ❸



For peak detection, here shown for 1D H-NMR, data we have developed a naive zero area filter (ZAF) which is fully data derived. The filter shape is sampled from a defined peak (ppm standard) and the ZAF has one parameter - the noise level, which is easily sampled and only used for sieving out small false positive peaks. For X/MS the ZAF is used to peak detect multimodal PICs.

Histogram Matching/normalization ❹



Histogram Matching normalization ❹ is based on a location-free approach to internal standard-free normalization. The method computes the multiplicative factor that makes the spectral intensity histograms as similar as possible. The HM method also works for TICs but usually X/MS data has multiple internal standards. The Histogram Matching does not have any critical parameters.

Benchmark

For comparison of the TracMass algorithm we have compared it with two other software, one academic - XCMS [1] and one commercial. The XCMS was compared using a LC/MS dataset that was published along with the XCMS algorithm/method. It comprises LC/MS data from a study of 20 IS spiked human serum samples designed to mimic a biomarker situation. Here TracMass found 10200 peaks vs. 2700 for XCMS (100% consistency). TracMass made a 106% recovery of ISes compared to XCMS 94%.

The commercial software is compared using a dataset of 99 UPLC/MS-QC runs (over time) of one pooled plasma sample. Here TracMass revealed 7000 peaks (100% consistency) and the commercial software 2100 (90% consistency).

For 1D-NMR (600MHz) of urine, the ZAF detection typically reveal 1200 peaks/spectrum. A standard GFHT run aligns about 800 peaks - the rest are spurious and/or false-positives. This is to be compared with the golden standard method of (0.04ppm) bucketing resulting in <256 buckets. Furthermore, the information carried in the GFHT aligned data is not confounded with other signals.

To conclude: the proposed processing pipeline can reveal more features in the data (ZAF & TracMass)- more than a factor of three. The GFHT can successfullt align these data. The result is put in tabular format. The processing pipeline is fast and delivers data with supreme confidence.

[1] Smith C.A., Want E.J., O'Maille G., Abagyan R., Siuzdak G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification, Analytical Chemistry, 2006, 78, 779-787.

References

❶ Åberg KM, Torgrip RJO, Kolmert J, Schuppe-Koistinen I, Lindberg J:
Feature detection and alignment of hyphenated chromatographic-mass spectrometric data. Extraction of Pure Ion Chromatograms using Kalman tracking. J Chromatogr A 2008, 1192(1):139-146

❷ Alm E, Torgrip R, Åberg K, Schuppe-Koistinen I, Lindberg J:
A solution to the 1D NMR alignment problem using an extended generalized fuzzy Hough transform and mode support. Anal Bioanal Chem 2009, 395(1):213-223

❸ Csenki L, Alm E, Torgrip R, Åberg K, Nord L, Schuppe-Koistinen I, Lindberg J:
Proof of principle of a generalized fuzzy Hough transform approach to peak alignment of one-dimensional 1H NMR data. Anal Bioanal Chem 2007, 389(3):875-885

❹ Torgrip R, Åberg K, Alm E, Schuppe-Koistinen I, Lindberg J: A note on normalization of biofluid 1D 1H-NMR data. Metabolomics 2008, 4(2):114-121

❺ Åberg K, Alm E, Torgrip R: The correspondence problem for metabonomics datasets. Anal Bioanal Chem 2009, 394(1):151-162