The "Mechanical Mass Spectrometrist", or: How to Collect 100,000 Manual Peak Annotations

Manor Askenazi[1], Hisham Ben Hamidane[2], Anja M. Billing[2], Rudolf Engelke[2], Sara E. Lendal[2], Sunkyu Choi[2], Johannes Graumann[2,*]

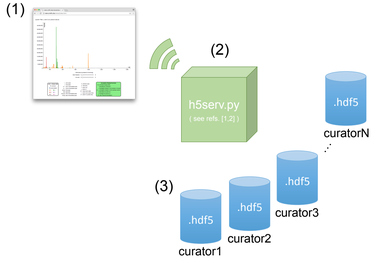
[1]Biomedical Hosting LLC, Arlington, MA, [2]Weill Cornell Medicine - Qatar, Doha, Qatar *Currently: Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany

Overview

In this work we present:

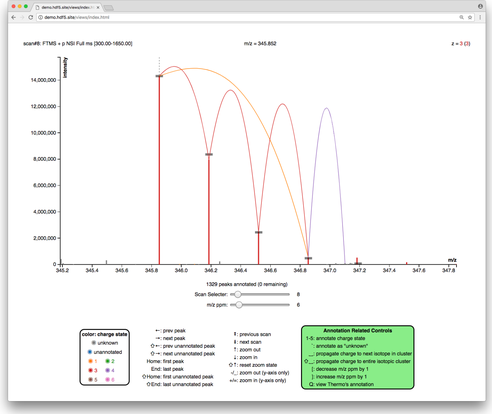
1) A web-based annotation tool for interactive charge state calling,
2) An HDF5-based solution for server-side storage, query and update of the resulting annotations,
3) A Gold Standard, manually generated library of 108,012 charge state calls.

The purpose of the work is to support the development of mass informatics by facilitating the manual curation of "Gold Standard" datasets to be used by both Human algorithm designers as well as AI/ML systems.



Introduction

Manually annotated datasets constitute an essential pre-requisite for both rational algorithm design and machine learning techniques. For certain tasks, (e.g. simple image recognition), non-specialists can provide valuable annotations using standard tools, e.g. through services such as **Amazon's "Mechanical Turk"** (http://www.mturk.com). However, for highly technical tasks arising in mass spectrometry, reference annotations must be collected from skilled users through custom curation tools. We have developed such a tool targeting the task of charge state deduction and employed it to thoroughly document over 100,000 QExactive peak-curation events. The resulting dataset constitutes the largest manually curated gold standard for charge state elucidation available to date.



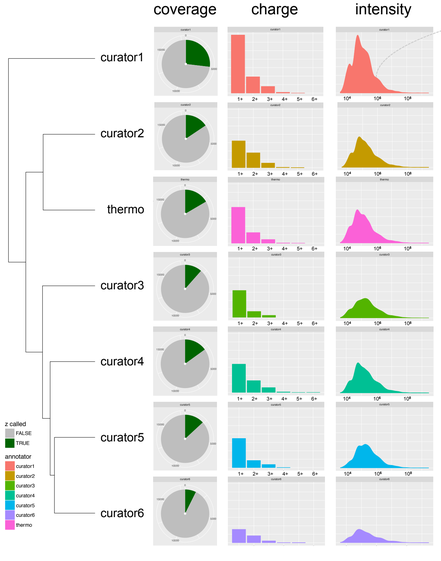
A Browser-Based Curation Tool was implemented in HTML5 using JavaScript-driven interactive visualizations of mass spectral peaks rendered in SVG. During the annotation process and specifically upon focusing on a given peak (e.g. m/z = 345.852, in this case), annotators were provided with three forms of additional supporting information. Specifically, annotators were: (i) shown all peaks pairs with m/z distances consistent with charge states one through six (in the form of color-coded arcs as seen in above), (ii) able to call up a predicted isotope distribution assuming the peak being studied was the monoisotopic peak, (iii) given access to the "solution" recorded by Thermo's data-acquisition software at run-time. Note that the annotation tool requires users to actively assign the "unknown" ("?") category, thus ensuring that spectra are annotated to completion.

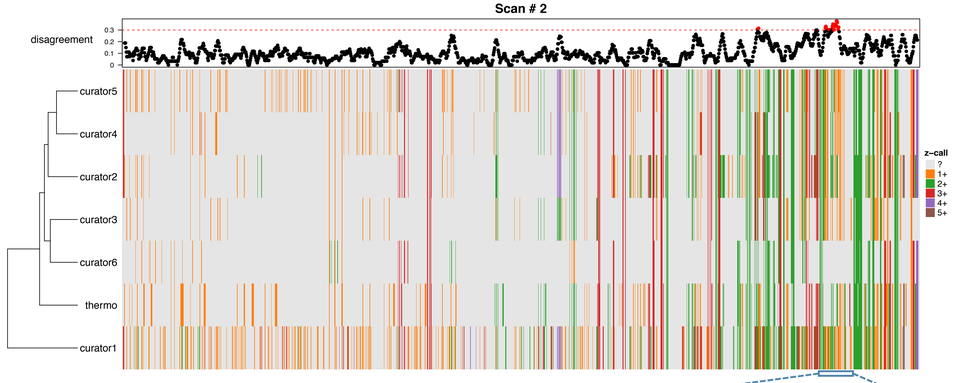
A pure HDF5-based server-side storage format was selected for both the data being visualized as well as user annotations. A pre-existing HDF5 RESTful API server (h5serv) implemented by the HDFGroup was used [1,2], essentially eliminating the need for custom server-code, including any code necessary for user-annotation write-back. An instance of h5serv was set up as a shared annotation system and six annotators were recruited: while these annotators spanned a relatively wide range in terms of mass spectrometry experience, all were sufficiently versed in the physical chemistry of mass spectrometry to recognize and label isotope clusters corresponding to a specific charge state of an unknown analyte. Each annotator was tasked with the comprehensive annotation of ten precursor ion/survey spectra selected at random from an LC-MS2 analysis of a commercially available tryptic digest of bovine serum albumin run on a QExactive. The spectra contained on average ~1800 peaks per spectrum and the exercise led to 108012 individually annotated peaks

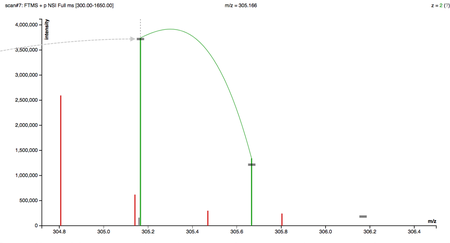
Results

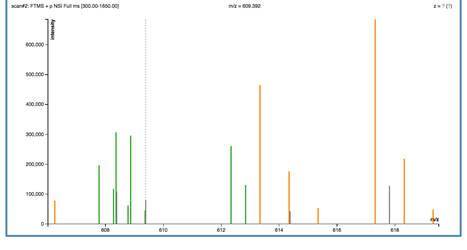
A clustering of the curator's work (shown below) was generated using a straightforward distance measure (pairwise disagreement rate). This clustering shows that, while the curators were all able to make charge state calls for a fair number of peaks (average calls made, i.e. coverage: 2729 or approximately 15% of the peaks) there was nonetheless wide variation both in terms of absolute number of calls (e.g. curator1 issued 4828 calls, which corresponds to 27% of data – as compared to, e.g. Thermo's acquisition software which recorded 16.5% non-zero charge state calls) as well as the distribution of charge states the curators tended to call and the intensity of peaks that they were willing to annotate with a non-zero call (zero indicates the inability to call a specific charge state).

Characterizing (Dis)Agreement amongst Curators is necessary prior to the generation of a final reference Gold Standard. For example, while curator1 is often in the minority (as the issuer of significantly more charge state calls), this may be either due to greater experience in mass spectrometry or a willingness to make unwarranted calls… In addition, it is important to verify that curators were not overly biased by their ability to inspect the annotations made by the Thermo acquisition software (since the web-based curation tool provided this information at curation time).

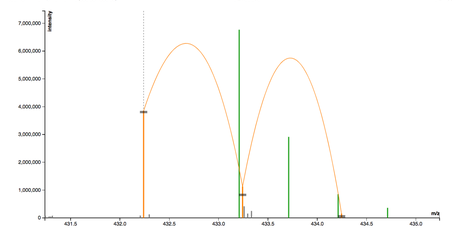








Standing Alone… (above) example annotation made by curator1, which was uniquely willing (among humans and machines alike) to call the rather intense peak at m/z = 305.166 as a charge state 2+! Conversely, (below) one can see a peak annotated unanimously by all the human curators as being a charge state 1+ peak, whereas Thermo's acquisition software assigned it "unknown" status…



Beware the "danger zone"!!! Seen above is a region of scan #2 exhibiting a relatively high disagreement rate across all curators (where disagreement is measured as the proportion of all pairwise disagreements per peak). The resulting smoothed disagreement rate is shown above the clustering (after application of a moving average filter with a symmetric 11 point window). Flagging such regions can be useful as part of the post-processing done prior to the release of a final Gold Standard dataset.

Updates about this project can be found at:

http://blog.biomedical.hosting/mechanical-mass-spectrometrist

References

[1] *The arc of Mass Spectrometry Exchange Formats is long, but it bends toward HDF5.* Manor Askenazi, Hisham Ben Hamidane, Johannes Graumann, **Mass Spectrometry Rev**, 2016 Oct.14 [Epub ahead of print]

[2] https://github.com/HDFGroup/h5serv