HDF5 File Architecture for Efficient LCMS Data Acquisition, Archival and Access

Jeff Jones, Ryan Benz SoCAL BIOINFORMATICS INC, Montrose CA

www.socalbioinformatics.com

Summary

FORMAT

HDF5 is a well developed and extensible big data format.

Structures exist for metadata strings, single dimension arrays, tables and matrices.

EFFICIENCY

HDF5 minimizes disk space requirements by nearly a factor of 2 from vendor formats, and over 5 times better than XML based mzML.

HDF5 allows access to portions of the data without loading the entire file to memory, called a *hyperslab*.

EXAMPLES

Recomputing the volume under the isotope across 100s of experiments for a given feature or missing feature of interest at a known LC MZ.

Searching for similar ms2 spectra for a given feature of interest at a known LC MZ to perform spectra addition to improve peptide sequence identification.

MOTIVATION

An extensible data format for large reproducible studies involving long duration clinical trials or longitudinal monitoring via liquid biopsies.

Surprisingly we record mz along with intensity for every spectra, despite mz rarely changing, resulting a massive data redundancy.

XML is terribly space inefficient, and while mz5 is a decent attempt at the HDF5 format it lacks an implementation to extract data portions without loading the entire file to memory. We contend that the simpler the data structure, the more accessible it will be to researchers and ubiquitous computing technologies, such as R or Python data-tables and SQL. This also facilitates the distributed cloud architectures such as AWS, Azure, Digital Ocean or local implementations with SMS, RabbitMQ or Celery.

While there is a clear motivation to adhere to the HUPO Proteomics Standards for data integrity, there are only a few essential data constructs:

- Time-Stamped Measurements
- Measurement Conditions
- M/Z ~ Intensity

In addition to having a basic standard approach, HFD5's architecture allows project unique data-constructs that facilitate research exchange and be attached to the data rather than existing in a referenced database or file.

DESIGN

Combination of SQL-like relational tables with numeric arrays and matrices

RAW DATA

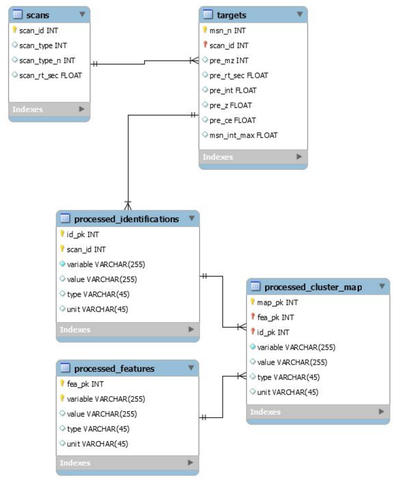
MSn intensity as a **matrix**, mz as **array**

META DATA

MSn reference as **data table**

PROCESSED DATA

MS1 Features as a **data table**
Peptide IDs as a **data table**
Cluster Mapping as **data table**



FUTURE

Vendor adoption of an HDF5 file format would facilitate efficient disk management, open the file for addon data while continuing support for simultaneous read/write functions that enable vendors to display real-time spectra and metrics in addition to opening the possibility for near-real-time biological identification. We also anticipate an enormous potential for DIA data efficiencies which capture both ms1 and ms1+n in the same fashion. We currently employ this proposed data format for use in large data sets where computations involving multi-sample and longitudinal studies are required.

Find an R implementation on Github @ https://github.com/jeffsocal/mzh5

EXAMPLE

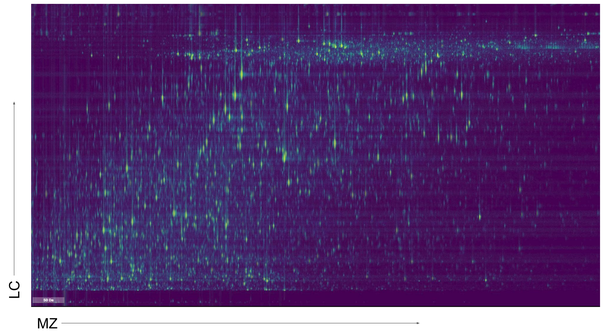
SCIEX TripleTOF 5600

file	size	relative
*.wiff	0.22 GB	-
mzML	1.95 GB	886 %
mz5	0.46 GB	210 %
HDF5	0.14 GB	64 %

AGILENT 6550 QTOF

file	size	relative
*.d	0.75 GB	-
mzML	3.67 GB	489 %
mz5	1.43 GB	191 %
HDF5	0.47 GB	63 %

Collected data from 20 LCMS experiments across 2 vendors using digested whole plasma 30min LC gradient, DDA ms2.



DATA STRUCTURE (example demonstrated here accessed hdf5 data via standard R packages; h5, plyr, ggplot2)

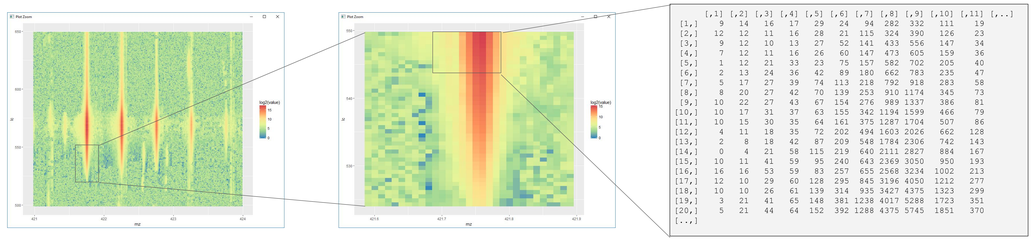
SCANS

scan_id	scan_type	scan_type_n	scan_rt_sec
200	2	34	81.43
201	2	35	81.53
202	2	36	81.63
203	2	37	81.73
204	1	167	82.15
205	2	38	82.45
206	2	39	82.55
207	2	40	82.65
208	2	41	82.75
209	1	168	83.18

TARGETS

msn_n	scan_id	pre_mz	pre_rt_sec	pre_int	pre_z	pre_ce	msn_int_max
200	410	759.8197	127.02	33	2	7.2	458041.4
201	411	762.8189	127.12	35	2	20.0	795969.6
202	412	822.8033	127.22	49	2	8.4	118907.0
203	416	855.8030	129.11	49	2	13.0	884632.8
204	418	792.8393	130.12	39	2	9.7	413187.0
205	420	850.8339	131.13	22	2	16.0	899425.9
206	422	705.8772	132.15	36	2	13.0	632037.9
207	424	768.8353	132.87	44	2	14.4	332029.4
208	425	823.8497	132.97	30	2	6.4	137225.7
209	426	831.8055	133.07	23	2	9.6	76881.5

MSn intensity matrix



```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,..
 [1,]    9   14   16   17   29   24   94  282  332   111   19
 [2,]   12   12   11   16   28   21  115  324  390   126   23
 [3,]    9   12   10   13   27   52  141  433  556   147   34
 [4,]    7   12   11   16   26   60  147  473  605   159   36
 [5,]    1   12   21   33   23   75  157  582  702   205   40
 [6,]    2   13   24   36   42   89  180  662  783   235   47
 [7,]    5   17   27   39   74  113  218  792  918   283   58
 [8,]    8   20   27   42   70  139  253  910 1174   345   73
 [9,]   10   22   27   43   67  154  276  989 1337   386   81
[10,]   10   17   31   37   63  155  342 1194 1599   466   79
[11,]   10   15   30   35   64  161  375 1287 1704   507   86
[12,]    4   11   18   35   72  202  494 1603 2026   662  128
[13,]    2    8   18   42   87  209  548 1784 2306   742  143
[14,]    0    4   21   58  115  219  640 2111 2827   884  167
[15,]   10   11   41   59   95  240  643 2369 3050   950  193
[16,]   16   16   53   59   83  257  655 2568 3234  1002  213
[17,]   12   10   29   60  128  295  845 3196 4050  1212  277
[18,]   10   10   26   61  139  314  935 3427 4375  1323  299
[19,]    3   21   41   65  148  381 1238 4017 5280  1723  351
[20,]    5   21   44   64  152  392 1288 4375 5745  1851  370
[...]
```