

Wei Chen
November 2007

1. Lipidomics

The lipidomics dataset only records a list of peaks that correspond to some interested materials. It is basically a 1D data (array). The original data format is *.wiff. After the processing, the dataset is typically stored in an .xls file. For each identified material, it records the intensity. The comparison of the intensity lists among different (patient) samples is the main cancer diagnosis approach.

2. Genomics

The SNP data from genomics is represented with a sequence of symbols. It is basically a 1D data (array) and should have the similar data format as lipidomics. I haven't checked this dataset carefully yet.

3. Proteomics

Liquid chromatographic and mass spectrometry (LCMS) is the dataset for proteomics. It is basically a 2D data (array). It is time-varying data which records the intensity of different mass-spectrometry in a given solid space. Every data point has three attributes: time, mass, and intensity. A promising visualization technique is to render it as a terrain-like height field, of which the intensity is the height at a given (time, mass) location. One standard format of LCMS is mzXML, a standard file protocol. The reader for mzXML can be found in [MZXML.doc](#) and [mzXML 2.1 tutorial.pdf](#).

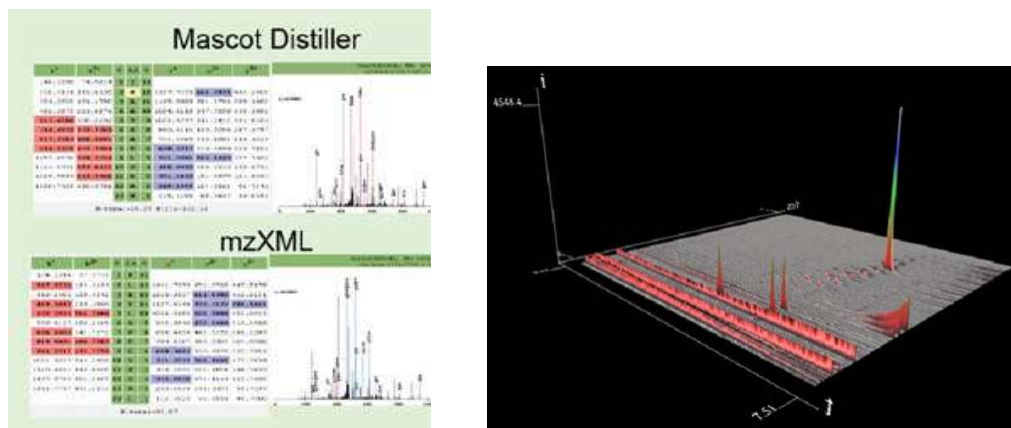


Figure 1. The LCMS is visualized as a 2D curve (left) or a 3D terrain (right).

4. Metabolomics Data

I have focused on the metabolomics dataset since this October. It is basically a 3D data (array). In terms of data dimensionality, other datasets such as proteomics, lipidomics can be regarded as a special case of metabolomics dataset. An essential data from *metabolomics* is the *GC (Gas chromatatographic) xGC Time-of-Flight (TOF) data*. In Dr.Rafetty's lab, the raw data acquired from an equipment is formatted as *.CDF (also NetCDF, binary). The propotol for netCDF can be found in

<http://www.angelfire.com/nt/jopearcy/ANDIandNetCDF.htm>. Typically, the users convert *.CDF into a text or xls file for further processing and analysis. The output file records the entire or a part of the 2D TIC (total ion count) image (not the whole 3D GCxGCxMass dataset). The TIC is the sum of intensities of all mass/z values for the sample. Currently, they also have interests to explore the metabolomics data with the detailed mass (intensity) data.

4.1 Data format of CDF

The structure for a *CDF is as follows:

- a) The global header which holds some global information of the dataset, including ndims (the number of dimensions), nvars (the number of all variables), ngatts (the number of global attributes) and recdim (The ID of the unlimited dimension, if there is one). Here, the variables correspond to the actual information. Table 2 list 18 variables readed from an example of Rafetty's lab:

Variable Name
error_log
a_d_sampling_rate
a_d_coaddition_factor
scan_acquisition_time
scan_duration
inter_scan_time
resolution
actual_scan_number
total_intensity
mass_range_min
mass_range_max
time_range_min
time_range_max
scan_index
point_count
flag_count
mass_values
intensity_values

- b) The detailed information about the dataset, including the variable ID (if any), the dimension name, the dimension size and values. Each dimension corresponds to a property of the handled GCxGC-MS dataset. For instance, the dataset I got from Rafetty's lab has 12 dimensions. And each dimension name and size is listed in Table 1:

Dimension Name	Dimension Size	Description
_2_byte_string	2	Unknown
_4_byte_string	4	Unknown
_8_byte_string	8	Unknown
_16_byte_string	16	Unknown
_32_byte_string	32	Unknown
_64_byte_string	64	Unknown

_128_byte_string	128	Unknown
_255_byte_string	255	Unknown
Range	2	Unknown
Point_number	112413155	The size of all valid samples (GC1, GC2, the intensity of each kind of mass)
Error_number	1	Unknown
Scan_number	216800	The multiplied size of all time samples (GC1, GC2)

When we have the Scan_number and Point_number, we can decide the actual size for GC1xGC2xMass. The dimensions of GC1 and GC2 are obtained from the equipment and is not recorded in *.CDF data. In Rafeytte's lab, for example, the dimension of GC1 is by default set as 400. Thus, the dimension of GC2 is $216800/400 = 542$. Because the mass values range from 50 to 900, the dimension for mass is 851. And thus the GC1xGC2xMass data has the dimension of (400x542x851). In other words, the dataset contain a 400x542x851array, of which each element records the intensity of the given GC1 (retina time 1), GC2 (retina time 2) with the given mass (ranges from 50 to 900).

c) The detailed information about all samples

Now we come to the real information. Each sample of the 400x542x851 dataset is recorded as a list of information, as shown in Table 3.

Variable Name	Example	Description
scan_number	- 9999.000000	Unknown
a_d_sampling_rate	50	Unknown
a_d_coaddition_factor	120.480000	The beginning time for GC1
scan_acquisition_time	0.010000	The beginning time for GC2
scan_duration	- 9999.000000	Unknown
inter_scan_time	- 9999.000000	Unknown
resolution	-9999	Unknown
actual_scan_number	-9999	Unknown
total_intensity	3594.000000	TIC
mass_range_min	50.000000	The minimum of mass
mass_range_max	900.000000	The maximum of mass
time_range_min	- 9999.000000	Unknown
time_range_max	- 9999.000000	Unknown
scan_index	29394	Unknown
point_count	600	Actual number of samples with valid intensity (i.e., intensity !=0). Thus, it could be smaller than 851.
flag_count	0	Unknown
mass_values	50.000	The mass value
intensity_values	18.000	The intensity of the mass value

All samples are listed sequentially in orders of the fields a_d_coaddition_factor (GC1), scan_acquisition_time (GC2) and mass_values (mass).

I wrote a toolkit to format the information loaded from *.CDF to a *.TXT file.

4.2 Data Analysis for GCxGC-Mass

In GCxGC-MS, each sample output of the GCxGC separation is analyzed by MS, e.g., time-of-flight (TOF) MS. Instead of a single value at each sample output, there is a mass spectrum consisting of an array of (mass/z, intensity) pairs. These mass spectra can be used in identifying unknown chemicals and in separating co-eluting peaks for more accurate quantification. Traditionally, a GCxGC TOF data is displayed and processed as a 2D image, of which each pixel represents the data sampled at (GC1xGC2) and its value is the sum (the total ion count, TIC) of all intensities for all sampled mass (mass per charge). **Figure 2** shows such an example. From the viewpoint of dimensionality, the TIC image is identical to a LCMS dataset and thus can also be visualized as height field. Most of current image processing, visualization, processing, and analysis tools for GCxGC-MS data are designed for GCxGC-TIC image.

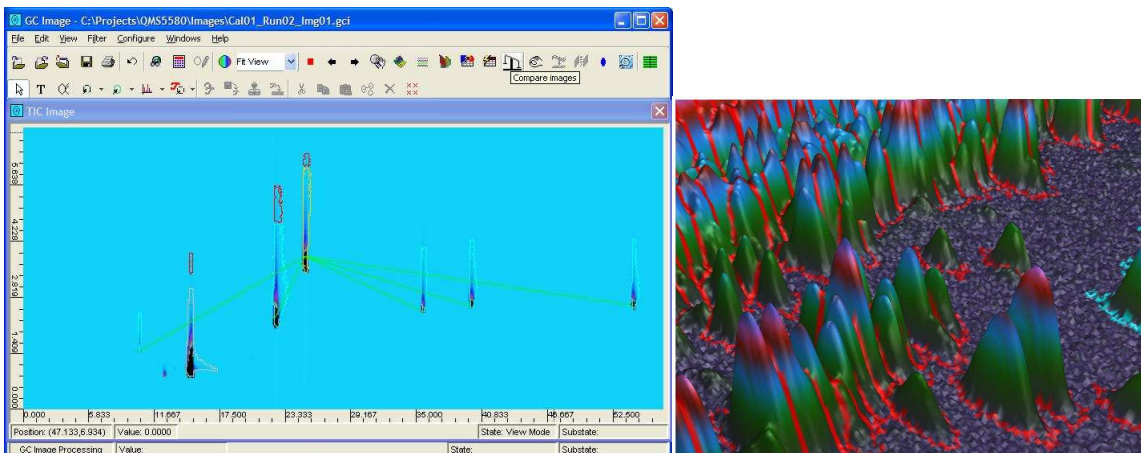


Figure 2: An image of the total ion count (TIC) values for GCxGC-MS data. From the viewpoint of dimensionality, the TIC image is identical to a LCMS dataset and thus can be visualized as height field too.

GCxGC-MS analyses generate very large data sets that may exceed the memory capacity of desktop systems, causing very slow processing. One approach to processing GCxGC-MS data more quickly is to reduce the size of the GCxGC-MS data by retaining only the elements with the largest intensity values in each mass spectrum and using sparse matrices to store the reduced mass spectra. For example, a GCxGC-MS image with 756x800 GCxGC samples and 215 intensity values in each mass spectrum requires nearly one gigabyte of memory (using four byte integers for the mass/z and intensity values) as shown in **Table 4**. If only the 32 elements with the largest intensity are stored for each mass spectrum, the memory requirement is reduced drastically to about 0.15 gigabytes as shown in **Table 5**.

Mass/Z values array	$756 \times 800 \times 215 \times 4 = 520,128,000$ Bytes	496.03 MB
Intensity values array	$756 \times 800 \times 215 \times 4 = 520,128,000$ Bytes	496.03 MB

Total memory required to store both the arrays:	992.06 MB


Table 4. Memory required to store the full mass spectrum for each sample in a GCxGC-MS image.

Mass/Z values array	756 x 800 x 32 x 4 = 77,414,400 Bytes	73.82 MB
Intensity values array	756 x 800 x 32 x 4 = 77,414,400 Bytes	73.82 MB
Total memory for both the arrays		147.64 MB

Table 5. Memory required to store the 32 values with the largest intensity for each mass spectrum in a GCxGC-MS image.

Of course, the gain in storage and processing efficiency must be weighed against the loss of data. **Figure 3** shows the reduction of a mass spectrum to retain only the values with the largest intensity in a sparse representation.

m/z	Intensity
31	1200
32	1250
33	1300
34	1350
35	200
36	1225
37	12300
38	100
39	40
40	700
41	500
42	35000
43	300
44	4000
45	550



m/z	Intensity
33	1300
34	1350
37	12300
42	35000
44	4000

Figure 3: An example of mass spectrum data reduction with sparse representation.

Note that, mass/z values may be integer or floating point. Some operations on mass spectra (such as NIST MS Library Search) require integer mass/z values. The user can set the threshold for rounding up fractional values to the next largest integer. Both graphical and tabular views of the selected mass spectrum are illustrated in Figure 4.

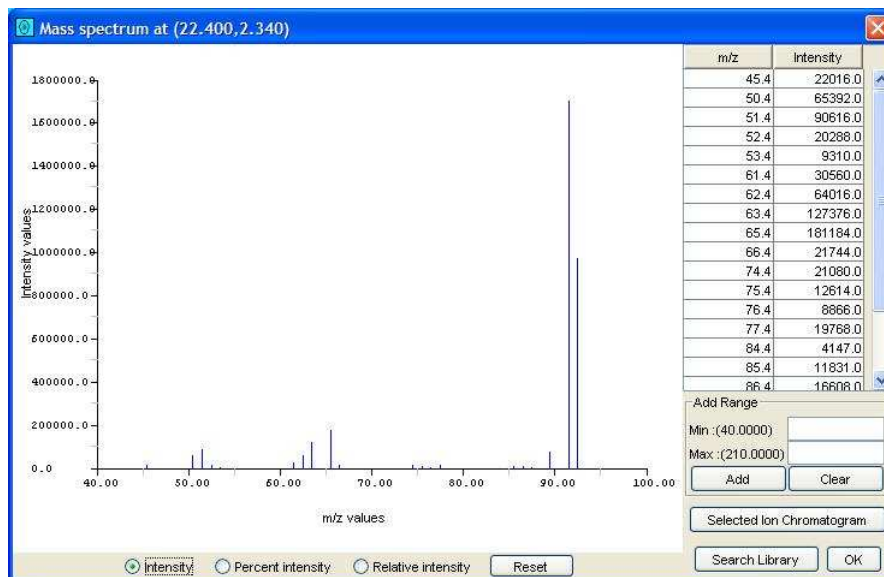


Figure 4: A simple viewer for the mass spectrum.

The tabular view (on the right side) has two columns: one for mass/z values and one for the corresponding intensity values in the mass spectrum. The intensity values can be expressed as absolute values, as percentages of total intensity (so that all values sum to 100), or as relative percentage values (so that the largest valued entry is 100.00). In the graph (on the left side of the window), the mass/z values are on the horizontal axis and the intensity values are on the vertical axis. The range for the horizontal axis is determined by the minimum and maximum mass/z values in all mass spectra of the image (so, the same absolute scale is used for the horizontal axis in all the mass spectra graphs). The domain for the vertical axis is from zero (or the minimum intensity if it is less than zero) to the largest intensity value in the selected mass spectrum. The vertical axis is expressed as intensity value, percentage of total intensity, or relative percentage. Sometime the users want to generate a selected ion chromatogram (SIC). A SIC is constructed with each pixel having the value of the total intensities in one or more mass/z subranges in the mass spectrum of the corresponding sample. An example SIC for the mass/z subrange 120-134 is illustrated in **Figure 5**.

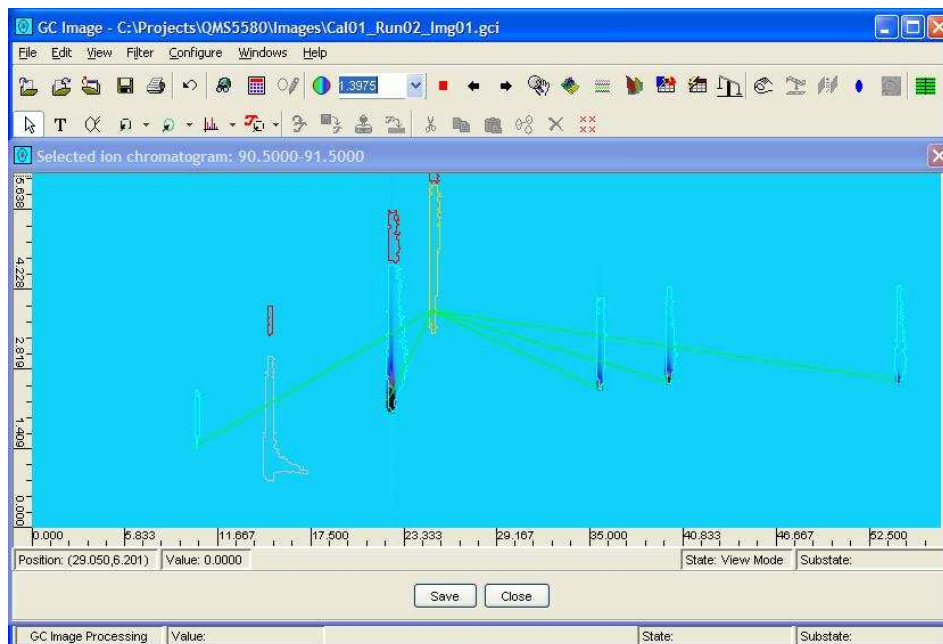


Figure 5: An example ion chromatogram image for the mass/z subrange 65-70.

With GCxGC Image, the basic operations are identity search and similarity search of the designated library (or libraries). If the unknown compound is likely to be in the library, identity search is the quickest way to find it. Identity search uses a very efficient method of selecting a small set of spectra for subsequent spectrum-by-spectrum comparison. The two types of identity search are Quick Search and Normal Search. Similarity search may be used for spectra that probably are not in the library.

The NIST Mass Spectral Search Program returns a "hit list" of matched chemical compounds from the library and several factors for each matched compound. They can be displayed in a table, as illustrated in **Figure 6**. Compounds in the hit list are listed with the following attributes:

- **Compound Name** is the primary name in the library. Synonyms are not listed.
- **Formula** is the molecular formula.
- The **Match Factor (MF)** is the normalized dot product with square-root scaling of the submitted mass spectrum and a library mass spectrum, using all the elements in the submitted mass spectrum.
- **Reverse Match Factor (RMF)** is the normalized dot product with square-root scaling of the submitted mass spectrum and the library mass spectrum, but the elements that are not present in the library mass spectrum are not included.
- **Probability** is the estimated relative likelihood of that the compound mass spectrum is the correct match for the submitted mass spectrum.
- **CAS Number** is the registry number of the matched compound in the library.
- **Molecular Weight** is the molecular weight of the matched compound.
- **Library** is the name of the library in which the matched compound was located.
- **Library ID** is the identification number of the matched compound in the library.
- **NIST#** is the NIST identification number of the matched compound.

Name	Formula	Match Factor	Reverse Ma...	Probability	CAS#	Molecular W...	Library	LibraryID	Select
Toluene	C7H8	928	929	36.61	108-88-3	92	mainlib	47426	<input type="checkbox"/>
1,3,5-Cyclohex...	C7H8	920	920	27.31	544-25-2	92	mainlib	47458	<input type="checkbox"/>
Spiro[2.4]heptane	C7H8	906	906	17.09	765-46-8	92	mainlib	47417	<input type="checkbox"/>
Spiro[3.3]heptane	C7H8	869	879	4.16	22635-78-5	92	mainlib	47414	<input type="checkbox"/>
1,6-Heptadiene	C7H8	861	875	3.1	5150-80-1	92	mainlib	47364	<input type="checkbox"/>
Cyclobutene	C7H8	860	870	2.98	52097-85-5	92	mainlib	47424	<input type="checkbox"/>
Tetracyclo[3.3.0.0.0]heptane	C7H8	846	846	1.86	278-06-8	92	mainlib	47461	<input type="checkbox"/>
2,5-Norbornene	C7H8	840	840	1.46	121-46-0	92	mainlib	47462	<input type="checkbox"/>
Benzeneace...	C8H9NO	829	830	1.0	103-81-1	135	mainlib	47435	<input type="checkbox"/>
1,5-Heptadiene	C7H8	827	827	0.92	3511-27-1	92	mainlib	47390	<input type="checkbox"/>

Figure 6: Example MS Search Hit List Table.