**Interim Report of LCMS Dataset Analysis**

*Pulasthi Wickramasinghe, Saliya Ekanayake, Geoffrey Fox*

*Indiana University, Digital Science Center*

July 11 2016

Analysis of Charge 2 peaks from dataset from Nature Article "Proteogenomics connects somatic mutations to signalling in breast cancer".

Internal run called “Stage 4”. Actual run details are identical to previous paper

Geoffrey Fox , D. R. Mani, Saumyadipta Pyne [Parallel Deterministic Annealing Clustering and its Application to LC-MS Data Analysis](http://dsc.soic.indiana.edu/publications/DAVS_IEEE.pdf) Proceedings of 2013 IEEE International [Conference](http://www.ischool.drexel.edu/bigdata/bigdata2013/index.htm) on Big Data October 6-9, 2013, Santa Clara, CA, USA, Technical Report June 23 2013 [DOI](http://dx.doi.org/10.1109/BigData.2013.6691636)

The final analysis steps are changed and results visualized with WebPlotViz <https://spidal-gw.dsc.soic.indiana.edu/> . The code is changed from C# to Java and run off on “Juliet” cluster – typically using 6-12 24 core Haswell nodes with Infiniband connection

**Cluster and Point Counts**

10,909,055 Charge 2 peaks in 37 experiments

97684 Single peaks (in “Sponge”)

423,393 clusters with 2 or more peaks as members

343,129 clusters with 5 or more members

6,805 clusters with 112 or more members

318 clusters with more than 300 members (not plotted)

Clusters defined by

Δ2D(*x*) ≤ c2

Δ2D(*x*) = ((m/Z|cluster center – m/Z|*x* )/ δ(m/Z))2 + ((RT|cluster center – RT|*x* )/ δ(RT))2

δ(m/Z) = D(m/Z) . (m/Z) and δ(RT) = D(RT)

Old Data DarTB D(m/Z) = 5.98 10-6 and D(RT) = 2.35 and c = 1 2 or 3 (or none)

New Data D(m/Z) = 2.17 10-6 and D(RT) = 3.13 and c = 2

**Number of Points in each experiment**

From Stage 4 fit, ratio of Number of Sponge (isolated single point clusters) points in each experiment to total number in experiment. Note average is 97684/10909055 = 0.0090 i.e. < 1%. All experiments have a small ratio with largest twice average

**m/Z Shifts**

Expectation value and standard deviation of

(m/Z|cluster center – m/Z|*x* )/ δ(m/Z)

As a function of experiment number

Most shifts are small

**RT Shifts**

Expectation value and standard deviation of

 (RT|cluster center – RT|*x* )/ δ(RT)

As a function of experiment number

Some Shifts are quite large

**Histograms of Peak Positions wrt centers**

This analysis uses 184614 Clusters with **occupation count cut Greater Than 20**. Histograms of

(m/Z|cluster center – m/Z|*x* )/ δ(m/Z)

And (RT|cluster center – RT|*x* )/ δ(RT)

Also shown is a Gaussian distribution of standard deviation 0.68. This implies that for this analysis true experiment standard deviation given as approximately 0.68 times δ(m/Z) and δ(RT). This implies cut of c times δ(m/Z) and δ(RT) is c/0.68 times standard deviations i.e. c=2/0.68 = 2.94 (near target of 3) times standard deviation.

**Visualization of Results** using WebPlotViz <https://spidal-gw.dsc.soic.indiana.edu/> and described in

*Supun Kamburugamuve, Pulasthi Wickramasinghe, Saliya Ekanayake, Chathuri Wimalasena, Milinda Pathirage, Geoffrey Fox, "TSmap3D: Browser Visualization of High Dimensional Time Series Data", Technical report May 10 201* [*http://dsc.soic.indiana.edu/publications/tsmap3d.pdf*](http://dsc.soic.indiana.edu/publications/tsmap3d.pdf)

Current browser technology cannot easily handle 10.9 million points and so we broke file for visualization into 20 separate sections sorted in m/Z. Further the m/Z precision is slightly truncated at moment; this will be corrected

All visualizations use x = ln(m/Z) / D(m/Z) and y = RT / D(RT), so error in same in x and y and uniform across plot. y (RT) is vertical axis and x (m/Z) horizontal. If you access web sites, please use Shift 2 to force 2D views (this is a 3D browser). To take a snapshot use Alt P. There are controls at bottom right to change sizes and speeds of zooming and panning.

There are two visualizations of the first (lowest m/Z values) section with ~half a million points and 20,000 clusters. Points (peaks) and determined cluster centers are displayed. Both visualizations display sponge points as orange stars and cluster centers as yellow circles.

<https://spidal-gw.dsc.soic.indiana.edu/resultsets/1074904939> has peaks categorized by cluster centers, sponge points and the particular one of the 423393 clusters points assigned to. In this case all points near a center will have the same color as they belong to same cluster with this center. Here you can press cluster button but it’s not so useful to scan 423393 clusters.

<https://spidal-gw.dsc.soic.indiana.edu/resultsets/748760675> has peaks arranged into 39 categories (clusters): Cluster centers, Sponge points, 37 experiments for non-Sponge points. Now each cluster is spread over entire dataset. Here the cluster list button bottom left is very useful to study individual experiments.

Note both these visualizations have identical points – they differ in labelling of points.

Above we present result of zooming out showing that points are arranged in clumps with few points (often sponge) between each clump. This is characteristic of dataset which has ~1600 clumps with just 30 in this little picture. Each clump has ~250 clusters and 6500 points

We now look in detail at just one – hopefully typical of these clumps.

The above shows one clump colored by experiment. This doesn’t show so much as all colors are present everywhere.

Here we select just 4 experiments and show fewer points. You need to zoom out the view but you can see interesting structure. There are clusters with expected 0-3 members but also some with many more.

First zoom in and look at top left. There are 3 “clusters”. The data for top two clusters, is in a line of roughly fixed RT. Presumably some anomaly?

Now look at structure near top of clump showing just 7 experiments colored distinctly. We see again clusters with expected 0-3 members but also some with many more.

Finally, below look at the whole clump colored by cluster. You see as expected all cluster centers surrounded by points of a common color. We that in general clustering and assignment of sponge points is clean and as expected. There are on bottom – two thirds over from left to right – two sponge points near each. These should be combined as a cluster of size two. This can be achieved with a follow up run looking at sponge points and see if they have clusters. We would do this for a “final analysis” but stopped here to capture key features of analysis.

