Computational Tools for Integrative Analysis of Array Data for Cancer Research

Sanjay Agravat
IBS574

April 20, 2012

Abstract

Cancer is caused by genomic alterations that disrupt normal cell function related to the regulation of DNA repair, genome stability, cell proliferation, cell death, adhesion, angiogenesis, invasion, and metastasis. These aberrations can be naturally caused by changes in genome copy number (through amplication, deletion, chromosome loss, or duplication), changes in gene and chromosome structure, gene expression, and point mutations. In this paper, I demonstrate an approach using High Performance Computing and Parallel Programming to analyze large amounts of data related to gene expression and copy number in tumor samples and describe an algorithm that extends GISTIC to filter genes help to identify candidate "driver" alterations in cancer. The new algorithm reduces the genomic regions of interest that GISTIC identifies and contains virtually the same set of oncogenes as has been identified in the literature by TCGA Gene Rank Lists.

1 Introduction

Diffuse infiltrative gliomas are by far the most common primary brain tumors in adults, especially its most malignant form, glioblastoma multiforme (GBM). Gliomas are a tumor that starts in brain or spine and arises from glial cells (non-neuronal cells. It is rarely curable with a high grade survival approximately 1 year and low grade has median survival of 11 to 16 years. In this paper, I use some Glioma datasets to validate my approach and algorithm to help identify candidate "driver" alterations involved in tumor genesis.

Glioma datasets are available from various providers including Rembrandt, The Cancer Geneome Atlas (TCGA), and the Vasari Feature Set. I obtained the data sets for this experiment from TCGA, using the Affymetrix HT_HG U133A Gene Expression Level 2 data and MKSCC CGH Level 3 Copy Number data.

1.1 Related Work

Other related work in this area includes GISTIC, RAE, and CONNEXIC. GISTIC looks at signficant alterations in the amplitude and frequence of CNA data. RAE takes a fine grained approach to peak detection and signal extraction. CONNEXIC uses gene expression and copy number data to identify regions of interest. The CONNEXIC approach highlights the fact that there are limitations to analytical approaches based on CNA data alone since CNA regions are typically large and contain

many genes, most of which are passengers that are indistinguishable in copy number from the drivers.

2 Methods

The Gene Expression data set is approximately 385x20,000 elements and the CGH data is approximately 445x225,000 elements. In order to construct a valid comparison, I had to use the same patients and the same genes in both data sets, I filtered the data set down to 383 x 15,000 and 383 x 78,000 elements for the gene expression and CGH data, respectively.

I developed an implementation of GISTIC based on the algorithm description in [2]. The program outputs gene lists that meet the test for significance for amplified and deleted regions. There are various threshold values that can be used as input to the algorithm including threshold for amplification, deletion, and q-values.

My extension of GISTIC involves generating a Pearson correlation matrix of the Expression data against the CGH data. Generating a correlation matrix of these two original data sets would have produced a 20,000 x 225,000 element matrix which is about 4.5 billion points and takes up 35GB of space on disk. However, since I pre-processed the data to use the same patients and matching gene probe symbols, the matrix was reduced in size to approximately 15,000 x 78,000 elements, or 1.5 billion elements (roughly 12GB of data).

I applied an algorithm to filter genes based on thresholds passed into my algorithm and validated the results against the TCGA Gene Ranker lists. The parameters I used for amplification were .35 for average CN, 2.0 for maximum CN; and for deletion I used -.01 for average copy number and -.75 for minimum copy number. Note that the maximum CN of 2.0 for amplification means there must exist a probe that has a CN of greater than 2.0. These parameters were chosen arbitrarily so further refinement for choosing these parameters needs to be explored.

2.1 HPC Technology and Tools

In order to do the analysis in an efficient manner for the data sets, I leverage the use of a High Performance Computing (HPC) platform. The platform is a Symmetric Multi Processor (SMP) cluster with 10 nodes and a total of 192 cores. I used a hybrid approach with Message Passing Interface (MPI) and OpenMP to distribute tasks and perform Single Instruction stream, Multiple Data stream (SIMD) operations. I used the Sun Grid Engine (SGE) to submit my processing jobs which uses a Portable Batch System to schedule jobs. Rather than reading from text files, I utilized HDF5 for Input/Output (I/O). HDF5 has the added advantage of supporting parallel I/O and a portable format that is compressed. I stored the file on a distributed file system known as NFS so that all nodes in the cluster could access the resource.

The data set was partitioned evenly (or approximately evenly) by the number of samples across the 10 nodes in the cluster. Each node in the cluster has either 16 or 24 cores (potential threads) that are available for processing each calculation for the pearson correlation coefficient. Each calculation can be done independently which allows for concurrent processing and more throughput for calculation. Figure 1 provides a visualization of how tasks can be partioned across various nodes.

The Pearson correlation coefficient is calculated by extracting the row vector for the gene expression data and the column vector for the CGH data and then performing the inner product cacluation for the correlation calculation using the standard score divided by the number of samples.

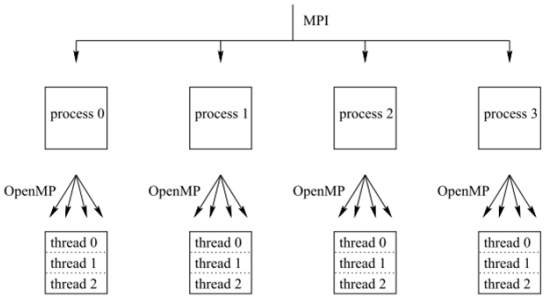

Figure 1: Task/Thread partitioning amongst various nodes using MPI and Open MP

$$GE_{n,t} \otimes CNV_{t,m} = \begin{bmatrix} \rho_{0,0} & \cdots & & \rho_{0,m} \\ \rho_{1,0} & & & \\ & \ddots & \ddots & \\ \rho_{n,0} & & \rho_{n,m-1} & \rho_{n,m} \end{bmatrix}_{n \times m} \tag{1}$$

$$\rho_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \tag{2}$$

2.2 Background Model

I created a background permutation model based on an all-to-all correlation of all genes to all probes for each patient. The test for signficance was done by calculating a z-score *in-situ* with a cutoff of 5.0 (pvalue of 1e-7) and $\rho = .248$. Elements that were greater than this cutoff were included as significant correlations. However, I did not do a False Discovery Rate correction but that will be included as future work.

2.3 Gene Filter Algorithm

For the purpose of this project, the significant correlations that I selected excluded different genes that correlated with each other. This is something I may consider revising in future work since there can be different genes that affect each other on the same pathway. After selecting the significant correlations, I run the list through my gene filter algorithm described below:

The algorithm takes a threshold for an average copy number and compares all the probe values for each patient to ensure that the threshold is met. It applies the same logic for the minimum and maximum threshold. The idea behind the average threshold is to ensure that we filter outliers where only small number of patients have an aberrant copy number. If the gene that it identifies is included in GISTIC then it adds it to the list; otherwise it discards it.

```
for all Significant Correlations do
    if type = del then
        if probe.avg < thresh and probe.min < minthresh then
            if probe.gene is GISTIC gene then
                rankedlist ← probe.gene
            end if
        end if
    else if type = amp then
        if probe.avg > thresh and probe.max > maxthresh then
            if probe.gene is GISTIC gene then
                rankedlist ← probe.gene
            end if
        end if
    end if
end for
```

Figure 2: Gene Filter Algorithm

3 Results

The performance of the parallel implementation for the Pearson Correlation matrix dramatically improved the performance of the calculation from 4 hours using 1 core to 12 mins using 192 cores (Figure 3). There are further areas of performance improvement that will be addressed as future work.

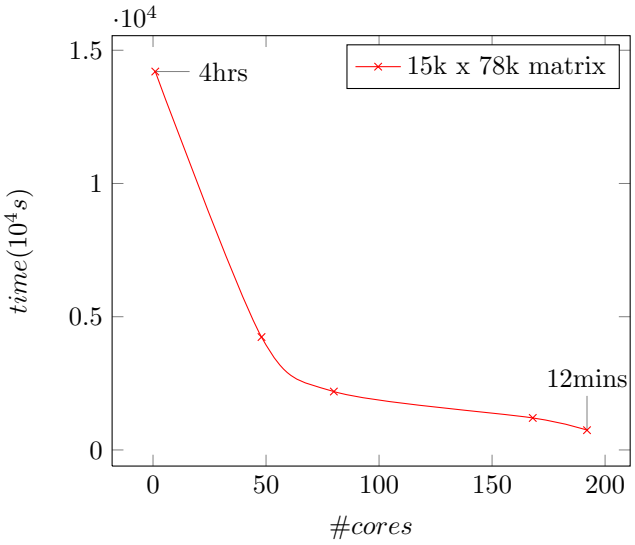

Figure 3: Performance Time vs. #Cores

3.1 Correlation Results

There were approximately 1,800 unique Probes from Gene Expression data that correlated to 49,000 probes from CGH array. One of the highest ones ($\rho = .71$) was the Copy Number for SLC38A1 (Chr 12) to the Gene Expression for CEP170 (Chr 1). The genes were Copy Neutral ($\mu = -0.01$, $\sigma = .37$) and were not highly expressed ($\mu = 13.5$, $\sigma = 1.5$). The standard deviation shows that there is not much variation in the values which supports the high correlation between these two genes. I also performed a pathway analysis using Ingenuity Pathway Analysis (Figure 4) and discovered that these two genes are in the same pathway. This further corroborates why the correlation was so high.

The correlation results from genes on the same chromosomes showed 2,500 unique probes from GE correlated to 13,500 unique probes. This higher degree of correlation is expected amongst genes o the same chromosome due to the way some genes are regulated in close physical proximity to each other.

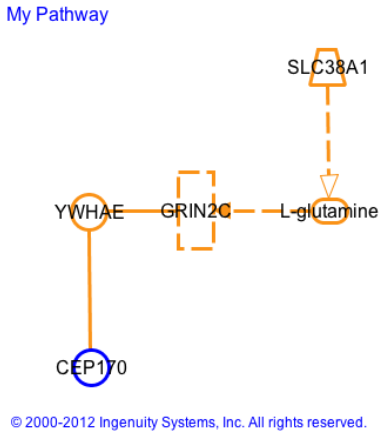

Figure 4: Pathway analaysis for CEP170 and SLC38A1

3.2 Gene Filter Algorithm Results

In order to validate the results of the Gene Filter Algorithm, I compared the results to the TCGA Gene Rank List which contains over 5,000 genes that have been indexed as significant genes by other algorithms such as RAE or from Literature references.

Many of the significant oncogenes that are tagged by TCGA Gene Rank were also found by the Gene Filter Algorithm; however, the Gene Filter Algorithm does not have a metric for ranking significant genes, so by default I used the pearson correlation coefficient. The Top 10-15 genes for amplification showed concordance for the oncogenes between the two methods, selecting genes such as EGFR, MET, and CDK4. However, the Top 10 genes for deletion showed a discordance since genes such as TP53, NIF1, and RB1 were not in the top 10. Those genes were in the lower 100-150 in my result list.

There were many genes that were found by GISTIC but not deemed significant by my correlation

method. The genes found by GISTIC were not included in my results; however, none of those genes found by GISTIC that were not signficant in my method were in the TCGA Gene Rank List. So it turned out to be safe to throw these genes and reduce the list using my approach.

These results indicate that the Pearson Correlation coefficient is not the ideal parameter to rank genes by. There is a potential to do some more research on combining parameters or trying to use other knowledge from pathway analysis or Gene Ontology function annotations.

gene	chr	avg	max	min	rho	generank	score	genelists
EGFR	7	2.03	4.61	−0.97	0.75	3	15.75	19
MET	7	0.4	6.05	−0.82	0.64	8	11.75	14
CDK4	12	0.43	4.4	−0.83	0.92	9	11.5	13
PDGFRA	4	0.34	0.58	−0.83	0.6	11	11	13
MDM2	12	0.32	0.62	−0.88	0.8	12	10.75	13
PDGFA	7	0.31	0.86	−0.83	0.3	67	6.5	8
TRRAP	7	0.35	0.88	−0.97	0.42	94	6	8
RHEB	7	0.33	0.88	−0.97	0.36	298	4	4
SMO	7	0.36	0.87	−0.64	0.32	345	3.75	5

Figure 5: Top 10 Amplified Regions Sorted by TCGA Gene Rank

gene	chr	avg	max	min	rho	generank	score	genelists
CDK4	12	0.43	4.4	−0.83	0.92	9	11.5	13
TSFM	12	0.33	0.62	−0.85	0.84	2,810	1.25	2
MDM2	12	0.32	0.62	−0.88	0.8	12	10.75	13
LANCL2	7	1.29	6.05	−0.82	0.79	2,646	1.25	2
GBAS	7	0.57	4.61	−0.97	0.78	2,153	1.5	3
KRIT1	7	0.38	5.12	−0.66	0.78	1,496	2	2
EGFR	7	2.03	4.61	−0.97	0.75	3	15.75	19
WIPI2	7	0.33	0.88	−0.97	0.72	2,828	1.25	2
IQCE	7	0.33	1.93	−0.81	0.7	4,161	1	1
RALA	7	0.35	0.88	−0.81	0.66	5,057	1	1
MET	7	0.4	6.05	−0.82	0.64	8	11.75	14
COPS6	7	0.35	0.88	−0.81	0.61	2,087	1.5	2
DDX56	7	0.34	0.88	−0.93	0.61	3,555	1	1
NUDCD3	7	0.33	0.88	−0.93	0.61	4,643	1	1
PDGFRA	4	0.34	0.58	−0.83	0.6	11	11	13

Figure 6: Top 14 Amplified Regions Sorted by ρ

gene	chr	avg	max	min	rho	generank	score	genelists
TP53	17	-0.06	0.55	−1.59	0.25	4	15.75	18
RB1	13	-0.21	0.83	−1.18	0.37	6	13.25	15
NF1	17	-0.02	0.28	−0.94	0.38	7	12.25	14
ATM	11	-0.05	0.52	−1.59	0.3	10	11.25	13
MLH1	3	-0.01	0.9	−0.91	0.42	24	8.75	11
SMAD4	18	-0.02	0.33	−0.94	0.33	26	8.75	12
ABL1	9	-0.02	0.33	−0.94	0.38	52	7	9
HSP90AA1	14	-0.13	0.62	−0.91	0.42	55	7	8
CTNNB1	3	-0.02	0.28	−0.94	0.38	57	6.75	9
HRAS	11	-0.08	0.21	−1	0.72	60	6.75	8

Figure 7: Top 10 Deleted Regions Sorted by TCGA Gene Rank

gene	chr	avg	max	min	rho	generank	score	genelists
MAPK14	6	-0.01	0.57	−1.92	0.76	616	3	5
HRAS	11	-0.08	0.21	−1	0.72	60	6.75	8
NDUFC2	11	-0.04	1.11	−0.75	0.72	4,545	1	1
STK38	6	-0.01	1.09	−1.32	0.72	5,386	1	1
PARK7	1	-0.12	0.66	−0.91	0.69	2,704	1.25	2
TFDP1	13	-0.12	0.55	−0.82	0.68	965	2.5	3
APTX	9	-0.15	0.42	−1.31	0.67	3,013	1	1
DIP2C	10	-0.42	0.84	−0.81	0.67	1,358	2	2
TBK1	12	-0.01	0.57	−1.04	0.66	963	2.5	4
KLHL9	9	-0.67	0.74	−0.8	0.64	4,237	1	1

Figure 8: Top 10 Deleted Regions Sorted by ρ

4 Concluding Remarks

Integrative analysis of Copy Number data with Gene Expression data produces encouraging results that validate established findings in the literature. However, the Pearson Correlation is a not a robust metric for correlation and future work should include a better metric such as the Spearman Rank Correlation. The strength of this measurement should also take into account the measure of entropy and variance in the data sets. A pair of data vectors that show high variance and high correlation is a stronger metric than a pair of vectors that have low variance but very high correlation.

While analyzing CGH data with Gene Expression showed some promise, it would also be potentially useful to perform methylation and expression comparisons. This analysis would help to illuminate cases where expression is low while copy number is high.

Utilizing Parallel Programming and High Performance Computing enables further exploration of Genomic data. As the deluge of genomic data increases, it will be necessary to leverage compute clusters or cloud resources to do the analysis.

References

[1] Uri David Akavia et al. "An Integrated Approach to Uncover Drivers of Cancer". In: *Cell* 143.6 (2010), pp. 1005 –1017. ISSN: 0092-8674. DOI: 10.1016/j.cell.2010.11.013. URL: http://www.sciencedirect.com/science/article/pii/S0092867410012936.

[2] Rameen Beroukhim et al. "Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma". In: *Proceedings of the National Academy of Sciences* 104.50 (2007), pp. 20007–20012. DOI: 10.1073/pnas.0710052104. eprint: http://www.pnas.org/content/104/50/20007.full.pdf+html. URL: http://www.pnas.org/content/104/50/20007.abstract.

[3] Barry S. Taylor et al. "Functional Copy-Number Alterations in Cancer". In: *PLoS ONE* 3.9 (Sept. 2008), e3179. DOI: 10.1371/journal.pone.0003179. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0003179.