Biomarker discovery via dependency analysis of multi-view functional genomics data

Ali Faisal
Helsinki Institute for Information Technology HIIT
Dept. of Information and Computer Science
Aalto University School of Science, Finland
ali.faisal@aalto.fi

Riku Louhimo
Computational Systems Biology Laboratory
Genome-scale Biology Research Program
University of Helsinki, Finland
riku.louhimo@helsinki.fi

Leo Lahti
Dept. of Veterinary Bioscience
University of Helsinki
Finland
leo.lahti@iki.fi

Sampsa Hautaniemi
Computational Systems Biology Laboratory
Genome-scale Biology Research Program
University of Helsinki, Finland
sampsa.hautaniemi@helsinki.fi

Samuel Kaski
Helsinki Institute for Information Technology HIIT
Aalto University and University of Helsinki, Finland
samuel.kaski@hiit.fi

Abstract

Cancers are complex diseases, characterized by genomic changes at multiple levels of regulation. We present an integrative genome-wide approach that captures shared patterns from several data sources and extracts chromosomal regions predictive of patient survival in glioblastoma multiforme (GBM) progression and drug resistance. Our results identify known and novel genomic regions that may contribute to GBM progression and drug resistance.

1 Introduction

In the recent decade cancer genomics has focused on the discovery of genetic mutations and chromosomal changes that support the cancer phenotype. Though a single mutation may relate to a particular phenotype, it is the combination of many different molecular mechanisms that disrupt cellular pathways and characterize a cancer [1]-[3]. A major effort in this context is the NIH's Cancer Genome Atlas project (TCGA) [4]. The aim of the consortium is to gather both multi-view molecular as well as clinical level characterization for patients in more than 20 different cancer subtypes.

Typically an integrative analysis is used to fuse and capture shared patterns from multiple data sources. There has been a considerable amount of research within the machine learning community on multi-view data analysis. In this work we extract shared patterns from multiple data sources using a machine learning model; specifically we use a Bayesian variant of constrained canonical correlation analysis (CCA). CCA is a multivariate statistical approach that detects linear dependencies by searching for their maximally correlated low dimensional representation. The Bayesian latent variable formulation not only allows us to model the noisy biological signals, but also provides a framework where sensible priors can be plugged in to encode specific relationships between different data sources.

The aim of this work is to identify potential regions (or biomarkers) that effectively stratify patients in low and high survival groups. We present an approach based on constrained

canonical correlation analysis that incorporates suitable priors, well-suited for multi-source integrative analysis in cancer genomics. In our study the clinical variable of interest is the number of days that respective patients survive. What is special in survival data is that there often are patients who survive over the entire study period and there are other patients with whom we lose contact. These observations that only contain partial information are termed censored data. *Survival analysis* methods handle censored data and in the study we use some of them to test survival association of chromosomal regions identified using the canonical correlation analysis.

Source code is available as an R package from: *http://research.ics.tkk.fi/mi/software/daSAr/*

2 Methods

2.1 Data

As a case study we selected glioblastoma multiforme (GBM) from TCGA. GBM is one of the most aggressive brain tumors; affected patients have a uniformly poor prognosis with median survival time of only 15 months over the past 25 years [5]. These tumors are now well characterized at genome and transcriptome levels and several studies have demonstrated that the combination of these two molecular levels may be advantageous to characterize robust signatures that are clinically relevant for GBM [6]-[7]. Three data types; gene-expression, DNA copy number changes and methylation pre-treatment measurements were collected for the available ~ 250 GBM samples. We used Anduril's *GetFromTcga* component that automatically downloads the latest version of the data from the TCGA database [8].

In the analysis we considered a chromosomal continuous data source as one view and gene-expression as the second view. This resulted in two studies: a) search for dependencies between copy-number and gene-expression, and between b) methylation and gene-expression. The probes for each dataset were matched resulting in ~3480 genes for gene-expression/copy-number pair and ~ 2530 genes for gene-expression/methylation pair. To satisfy the normality assumptions of our model, the data was log_2 transformed and the mean of signals for each probe was set to zero before the analysis. Besides the molecular profiles we included patient's clinical information such as age, gender and race in the analysis pipeline.

2.2 Dependency analysis

Unsupervised multi-view learning approaches are used to model multi-source datasets. We used a recently developed similarity-constrained canonical correlation analysis (simCCA) [9]. Canonical correlation analysis is an approach for capturing shared patterns from multiple views or data sources. It seeks a low dimensional transformation for the data sources such that the correlation in the latent space is maximized. The Bayesian generative formulation is as follows:

$$\mathbf{X} \sim N(\mathbf{W}_x\mathbf{Z}, \mathbf{\Psi}_x)$$
$$\mathbf{Y} \sim N(\mathbf{W}_y\mathbf{Z}, \mathbf{\Psi}_y)$$

where the two data sources $\mathbf{X}$ and $\mathbf{Y}$ are assumed to stem from the shared Gaussian latent variable $\mathbf{Z} \sim N(0, \mathbf{I})$ and normally distributed view-specific noise. The projection matrixes $\mathbf{W}_x$ and $\mathbf{W}_y$ encode the relationship between the data sources, and operate on the shared variable. Notice that in the classical CCA the projection matrixes operate on the original data (equations not shown). Correlation maximization of the classical CCA can be retrieved from the maximum likelihood solution of the Bayesian model [10]-[11]. The joint likelihood can be expressed as follows:

$$P(\mathbf{X},\mathbf{Y},\mathbf{W},\mathbf{\Psi}) = \int P(\mathbf{X},\mathbf{Y},\mathbf{W},\mathbf{\Psi})P(\mathbf{W}_y \mid \mathbf{W}_x)P(\mathbf{W}_y)P(\mathbf{\Psi})P(\mathbf{Z})d\mathbf{Z}$$

where $\mathbf{\Psi}$ denotes block diagonal matrix of $\mathbf{\Psi}_x$ and $\mathbf{\Psi}_y$. The conditional probability $P(\mathbf{W}_y \mid \mathbf{W}_x)$ encodes the relationship between the data sources $\mathbf{X}$ and $\mathbf{Y}$. It can be parameterized with a transformation matrix $\mathbf{T}$ such that $\mathbf{W}_x = \mathbf{T}\mathbf{W}_y$. This yields two extremes for the prior; in the unconstrained form, the approach reduces to the traditional CCA while setting $\mathbf{T} = \mathbf{I}$ yields identical shared components derived from both the data sources. SimCCA employs specific constraints via setting an appropriate prior on the transformation matrix and has been shown to outperform other learning methods in cancer genomics [9]. The prior on $\mathbf{T}$ can be used to make the model focus on searching for specific types of dependencies. We plug in a truncated normal distribution prior: $P(\mathbf{T}) = N_+(\|(\mathbf{T}-\mathbf{I})\| \,|0,\sigma^2\mathbf{I})$, where the variance parameter can be used to tune the tradeoff between the two extremes. For all other variables we use uninformative priors. For further details see [9].

Notice that this prior over $\mathbf{T}$ favors positive correlations among the two data sources, which is quite sensible in case of gene-expression and copy numbers. However for the case of gene-expression and methylation, the relationship is inverse; down-regulation of a gene can be due to hyper-methylation (an increase in the epigenetic methylation of cytosine and adenosine residues in DNA) and similarly up-regulation of a gene can be due to hypo-methylation. The inverse relationship is encoded by the prior $P(\mathbf{T}) = N_+(\|(\mathbf{T}+\mathbf{I})\| \,|0,\sigma^2\mathbf{I})$.

We implemented the model by defining a chromosomal region via a window that is centered at a gene and spans across ten neighboring genes within the chromosomal arm. The window was slid across all chromosomal arms and a dependency score and each sample's contribution towards the score for each region was calculated. The dependency score was computed as a ratio of strength of shared versus marginal affect: $Tr(\mathbf{W}\mathbf{W}^T)/Tr(\mathbf{\Psi})$, where Tr denotes matrix trace. A high score would reveal a correlating expression and corresponding chromosomal change; high-scoring regions with q-value < 0.05 were selected for further analysis. The significances of regions were estimated by a permutation test, using the observed dependency score as a test statistic. The samples in one of the spaces (gene-expression) were randomly rearranged removing the relationship with the other space (copy-number changes). One thousand such random permutations were formed and their dependency scores computed. Chromosomal region's significances were then determined as the proportion of random scores that were greater than the observed dependency score. For each identified region, sample-wise contribution scores were ordered and three groups were formed based on the 10th percentile, the 90th percentile and the rest. The same analysis was repeated for gene-expression and methylation dataset.

2.3 Survival analysis

In our study the variable of interest corresponds to death of the patient and we wanted to check if the stratified patient groups corresponding to each identified region had a significant survival association. Survival analysis approaches are commonly used to estimate the outcome variable of interest, namely the time until an event occurs. There are two main components in a survival analysis: estimation of survival function given censored data and comparison of the functions for multiple groups. The survival function $S(t)$ is the probability that an individual survives longer than time t. In our study we used the basic Kaplan-Meier (KM) estimator for the survival function given as

$$\hat{S}(t_{(j-1)}) = \hat{S}(t_{(j)})P(T > t_{(j)} | T \geq t_{(j)})$$

This gives the probability of surviving past the previous event time $t_{(j-1)}$, multiplied by the conditional probability of surviving past current time $t_{(j)}$, given survival to at least time $t_{(j)}$. The estimator allows us to draw KM survival curves for each group. The next step is to compare and test for their statistical equivalence. We use a log-rank test to compute significance for the differences [12].

Note that even though the KM analysis is used widely it does not model the effect of covariates, and hence the significance levels might be biased due to any confounding covariates. We checked for a bias using a Fisher contingency table analysis, where one of the groupings was induced by a quantile clustering on the sample-wise contribution scores from

the simCCA, and the second grouping was formed from any of the binary clinical variables considered separately. In the data we had three clinical factors; age, race and gender for each patient sample. These were transformed into the following binary variables; race: white/non-white, gender: male/female, and the age we discretized using four binary variables: age < 30, <40, >50 and >60 years.

An alternative model-driven approach to cater for external covariate is the Cox proportional hazard model. The Cox model is a regression-based approach which is an extension of the Kaplan-Meier analysis. It takes into account the effect of covariates on the given groups and adjusts the survival significances accordingly. Here we omit the details of the model for brevity and refer interested readers to references [12]-[13] for details.

The analysis was carried out independently for the gene-expression/copy-number data set, and the gene-expression/methylation data set. We searched for data-set specific as well as shared survival-associated regions, using both KM and Cox analysis.

3 Results

The dependency analysis resulted in 281 significantly dependent regions for the gene-expression/copy number datasets and 313 regions for the gene-expression/methylation datasets ($q < 0.05$). We observed that the histogram for patient contribution scores followed a bell shape centered at zero; that is, few patients contribute most to the dependency score; Figure 1 shows the histogram for four random regions from the gene-expression/copy-number datasets.

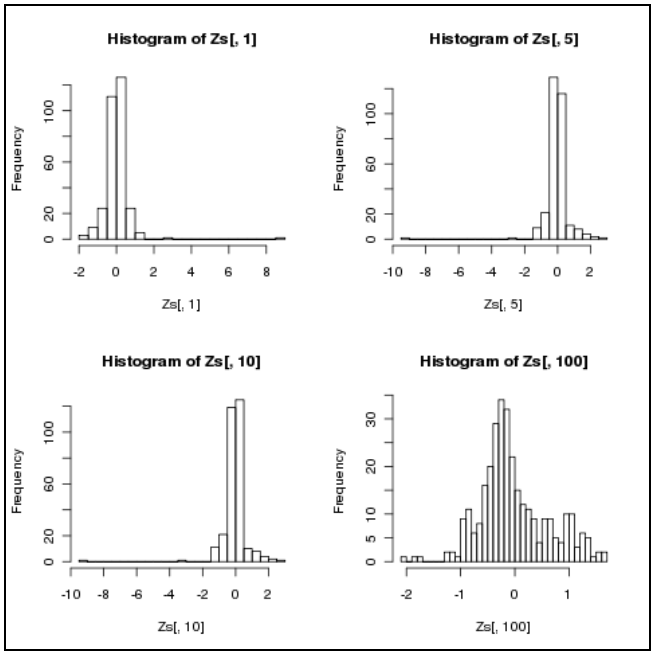

Figure 1: Histogram of patient contribution weights for four regions from the gene-expression/copy-number datasets. The vector Zs[,i] stores the contribution weights for a single region *i*. Samples that have a weights near zero have the least contribution to the dependency score.

The high positive and negative weights correspond to patients that contribute the most and we treat these as separate groups. The survival curve for each was compared independently with the group that did not contribute to the dependency.

From the gene-expression/copy-number datasets we found three significant chromosomal regions with a stringent cut-off on Kaplan-Meier survival association ($q < 0.05$): 10p13, 10q22.1, 10q26.13. Table 1 summarizes these regions. Many corresponding genes had expression profiles that correlated with copy number aberrations such as *HK1*, *HKDC1*, *MCM10*, *DDX21*, and *SLC29A3*. Figure 2 shows a sample KM curve for 10p13; other regions had similar curves. Copy number changes in chromosome 10 have been reported for brain tumors and specifically the 10q region has been shown to be closely related to glioblastoma [14]-[15]. The results from methylation/gene-expression revealed one statistically significant Kaplan-Meier-survival associated region ($q < 0.05$) at 21q22.2. This region, centered at *ETS2*, has both tumor suppressive and promoting properties depending on different tumor types [16]-[17].

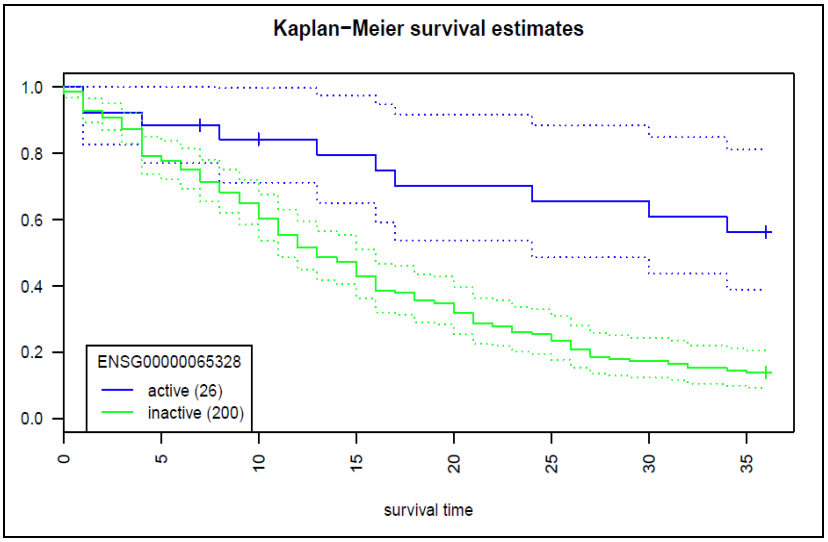

Figure 2: Kaplan-Meier survival curve for the region 10p13 centered at MCM10. Patients with a high dependency between copy number alteration and gene expression in the region ("active") have better survival association than patients having low dependency ("inactive"). X-axis: months, y-axis: percentage of GBM patients alive, round brackets: number of patients, dotted lines: 95% confidence intervals.

The Fisher enrichment test did not show any bias induced from the three external covariates; significance levels are reported in Table 1. Finally the integrative analysis of copy-number/gene-expression and methylation/gene-expression revealed one single shared significant region: 9p24.3 ($p < 0.05$) based on Cox analysis. Kaplan-Meier analysis did not find regions shared between the two data sets.

Table 1: Survival associated chromosomal regions identified using KM and Cox analysis from three different data types: gene-expression (exp), methylation (methy) and DNA copy number aberrations (cgh). Significantly enriched clinical factors are shown in red (qval < 0.05). A region is centered at the gene shown in bold.

														Enrichment Test (qvals)						
														Clinical factors						
	Dataset used	Biomarker	Genes in the region											White	Female	Male	Age < 30	Age < 40	Age > 50	Age > 60
KM	cgh - exp	10p13	**MCM10**	SEC61A2	OPTN	CDC123	OLAH	RPP38	PRPF18	PTER	CAMK1D	HSPA14		1	0.186444	1	0.254411	0.005766	1	1
		10q22.1	UNC5B	CHST3	SUPV3L1	HKDC1	**HK1**	DDX21	SGPL1	COL13A1	SLC29A3	KIAA1279		0.254411	0.378073	1	1	1	1	9.26078E-05
		10q22.1	HNRNPH3	UNC5B	SUPV3L1	**HKDC1**	HK1	DDX21	SGPL1	COL13A1	SLC29A3	KIAA1279		1	1	1	1	0.056428	1	1
		10q26.13	SEC23IP	**PLEKHA1**	BCCIP	WDR11	PTPRE	TACC2	BUB3	FAM175B	ACADSB	INPP5F		1	1	1	1	1	0.056428	0.811937969
	methy-exp	21q22.2	DOPEY2	**ETS2**	CBR1	MORC3	SLC37A1	PKNOX1	CRYAA	TTC3	MX2	SH3BGR		1	1	1	1	0.872796	1	1
Cox	cgh-methy & exp	9p24.3	**SMARCA2**	DNAJA1	TYRP1	SH3GL2	TEK	IFNA8	SNAPC3	NUDT2	KCNV2	PDCD1LG2		1	1	0.378073	1	1	1	1

4 Conclusion and future work

We use a combination of existing survival analysis techniques and a recent probabilistic machine learning approach to extract survival associated chromosomal regions from multi-view data sets. A case study on glioblastoma multiforme showed that with sufficient data, our approach indeed finds regions that are known to be actively involved and predictive of patient survival.

A related approach is genome-wide association analysis (GWAS) that searches the whole genome for small variations, called single nucleotide polymorphisms which occur more frequently in people with a particular disease than in people without the disease. The approach is however limited to a single data source. Our multi-view approach investigates the dependencies between different functional layers at the transcriptome and genome levels. This makes it possible to discover mechanisms and interactions that are not seen in the individual measurement sources.

The results highlight the need for advanced algorithms to define context at several levels in order to identify genomic regions or transcript profiles that play a key role in cancer progression and drug resistance. Typical analyses fall short in dealing with noise and uncertainty common in biomedical studies; we use a Bayesian dependency approach that circumvents these and incorporates a suitable prior well suited for multi-source analysis in functional genomics.

There are several fronts at which we can improve the analysis. A sensible extension is to develop the constrained CCA framework for more than two views. This basically involves redefining the cost function such that it maximizes dependencies among projections for all available data-sources. It is also meaningful to search for combinations of regions that together are better predictors of patient survival. Lastly, it would be very useful to have bigger targeted collections of homogenous patient samples e.g. administered with a specific drug, which would make it possible to study survival and other relevant variables of interest specific to a cancer treatment.

Acknowledgments

We would like to thank Vladimir Rogojin for helping us with Anduril and the data from TCGA. This work was supported by the Finnish Center of Excellence in Adaptive Informatics Research [AF, SaK], PASCAL2 Network of Excellence, ICT [216886] and Finnish Doctoral Program in Computational Sciences [doctoral fellowship to AF].

References

[1] Sherr CJ. (1996) Cancer cell cycles. *Science* **274**(5293):1672-1677.

[2] Hanahan D &Weinberg RA. (2000) The hallmarks of cancer. *Cell* **100**(1):57-70.

[3] Vogelstein B, Kinzler KW. (2004) Cancer genes and the pathways they control. *Nat. Med.* **10**:789-799

[4] McLendon R, et al., (2008) Cancer Genome Atlas Research Network: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**(7216):1061-1068

[5] Huse JT, Holland EC. (2010) Targeting brain cancer: advances in the molecular pathology of malignant glioma and medulloblastoma. *Nat. Rev. Cancer* **10**(5):319-31

[6] Nigro J.M, Misra A, Zhang L, Smirnov I, Colman H, Griffin C, Ozburn N, Chen M, Pan E, Koul D, Yung W.K.A, Feuerstein B.G, Aldape K.D. (2005) Integrated Array-Comparative Genomic Hybridization and Expression Array Profiles Identify Clinically Relevant Molecular Subtypes of Glioblastoma, *Cancer Research* **65** 1678-1686

[7] de Tayrac M, Etcheverry A, Aubry M, Saïkali S, Hamlat A, Quillien V, Le Treut A, Galibert MD, Mosser J. (2009) Integrative genome-wide analysis reveals a robust genomic glioblastoma signature associated with copy number driving changes in gene expression. *Genes Chromosomes Cancer* **48**(1): 55-68.

[8] Ovaska K, Laakso M, Haapa-Paananen S, Louhimo R, Chen P, Aittomäki V, Valo E, Núñez-Fontarnau J, Rantanen V, Karinen S, Nousiainen K, Lahesmaa-Korpinen AM, Miettinen M, Saarinen L, Kohonen P, Wu J, Westermarck J, Hautaniemi S. (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med.* **2**(9):65.

[9] Lahti L, Myllykangas S, Knuutila S, Kaski S. (2009) Dependency detection with similarity

constraints. *IEEE International Workshop on Machine Learning for Signal Processing*, pp 89-94. (Implementation available at: http://bioconductor.org/packages/release/bioc/html/pint.html)

[10] Bach F.R. and Jordan M.I. (2005) A probabilistic interpretation of canonical correlation analysis. *Tech. Rep. 688,* Department of Statistics, University of California, Berkeley.

[11] Klami A, Kaski S. (2007) Local Dependent Components. In Zoubin Ghahramani (Ed.), *Proceedings of the 24th International Conference on Machine Learning*, pp. 425-433. Omni Press

[12] Kleinbaum, D.G. and Klein, M., (2005) Survival Analysis. A self-learning text. Springer

[13] Bradburn M.J., Clark T.G., Love S.B. and Altman D.G. (2003) Survival Analysis Part II: Multivariate data analysis: an introduction to concepts and methods. *British Journal of Cancer* **89**(3): 431-436

[14] K Tokiyoshi, T Yoshimine, M Maruno, A K M G Muhammad, and T Hayakawa (1996) Accumulation of allelic losses on chromosome 10 in human gliomas at recurrence. *Clin Mol Pathol*. **49**(4): M218–M222.

[15] Rasheed B., Fuller G., Friedman A., Bigner D., Bigner S. (1992) Loss of heterozygosity for 10q loci in human gliomas. *Genes Chromosom Cancer* **5**(1):75-82.

[16] Sussan R.E., Yang A., Li F., Ostrowski M.C. and Reeves R.H. (2007) Trisomy represses Apc^{Min}-mediated tumours in mouse models of Down's syndrome. *Nature* **451**(7174):73-75

[17] Tynan J.A., Wen F., Muller W.J. and Oshima R.G. (2005) Ets2-dependent microenvironmental support of mouse mammary tumors. *Oncogene* **24**(46):6870–6876.