Integrated analysis of diverse genomic data

Georgia Tsiliki[1], Konstantinos Tsaramirsis[2], Sophia Kossida[1] ✉
[1]Biomedical Research Foundation of the Academy of Athens (BRFAA), Athens, Greece
[2]King's College London, London, United Kingdom

Motivation and Objectives

The increasing growth of high throughput genome-wide assays, such as next generation sequencing (NGS), is enabling the simultaneous measurement of several genomic features in the same biological samples. As a consequence forefront genome consortia have faced the challenge of integrating these diverse data types (Parson *et al.,* 2008; Network TCGAR, 2008), including RNA transcriptional levels, genotype variation, DNA copy number variation, and epigenetic marks. Often such data types produce controversial or partly overlapping results, towards a particular disease of interest, resulting in only a limited number of successful applications to everyday medical practice. Particularly in cancer research, this overall failure to translate modern advances in basic cancer biology is also attributed to the lack of comprehensively organizing and integrating all of the 'omics' features now technically acquirable on virtually any type of cancer (Vaske *et al.,* 2010). Two methodological approaches are mainly presented, namely meta-analysis techniques and integration techniques considering all the data types simultaneously, however until now there has not been developed a general and scalable statistical framework ready to incorporate as many diverse 'omics' features (Tyekucheva *et al.,* 2011). We present a model-based methodology, which considers all data together and aims in estimating important gene-sets for the pathology under study. An important objective is to validate established gene signatures, although emphasis is given on those sets which can only be found through integrated analysis. Special technical issues considered are different data formatting as well as data rescaling. The use case presented here considers breast cancer disease.

Methods

We consider microarray gene expression, RNA-seq and copy number variation measurements taken from the same samples as derived by The Cancer Genome Atlas (TCGA; https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp) database. An extra source of variation introduced in the data is due to the different technology platforms they are derived from. Particularly, the microarray gene expression data are derived from Agilent and Affymetrix platforms (AgillentG4502A _ 07 _ 03, U133A, U133Plus2), the expression mRNA sequencing data from Illumina platform (IlluminaGA _ RNASeq, IlluminaHiSeq _ RNASeq) and the copy number variants from Affymetrix platform (Genome _ Wide _ SNP _ 6.0). Nevertheless diverse data are taken from the same breast cancer samples, to avoid further discrepancies.

We present a Bayesian partition model to detect genetic interactions in the data, where a Markov Chain Monte Carlo (MCMC) algorithm is designed to simultaneously search across datasets (Denison *et al.,* 2002). The above methodology is a powerful modelling approach, which can handle large number of data, and also allow for interaction across data samples. Our aim is to present a stand-alone tool able to independently analyse the data using standard clustering methodologies (hierarchical clustering algorithms) and also provide the option of an integrated analysis for all data types simultaneously.

Towards this end, we first establish a common annotation mirror, where all entries are 'translated' to chromosomal locations. Data from only chromosomal regions across data sets are considered. The above procedure results in variable number of entries per chromosomal region, given the data set, for example a chromosomal location could include one gene and two copy number variants. This is addressed by averaging the data entries over the region. Similar averaging techniques have been applied to cross-platform microarray data in the past.

Regression modelling is employed for each chromosomal location and each sample, to test the null hypothesis that the particular chromosomal location is of no interest to breast cancer disease. The above procedure results in a single data set for all data types which is then further

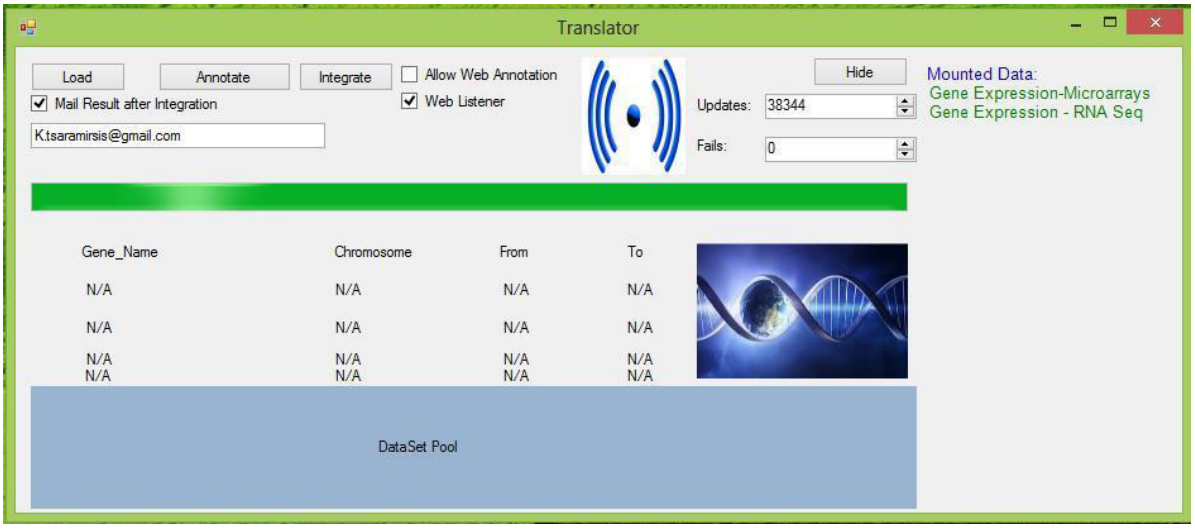

Figure 1. The integrated analysis stand-alone tool. The first form of the tool is shown, where users can upload the data and proceed with annotating or integrating the diverse data types.

analysed for estimating interesting gene signatures. The partition modelling is employed to the derived data set aiming to estimate clusters of homogeneous data which are then interpreted as breast cancer gene-signatures and cross validated with those derived from the individual data analysis and well-established gene-signatures.

In Figure 1, we show an instance of the application tool produced; users are prompted to upload the data sets one at a time, and can either proceed with 'annotating' the data using the chromosomal region scheme introduced above, or directly proceed with the integrated analysis.

Results and Discussion

Both simulated and empirical data examples demonstrated our method's ability to detect highly correlated data groups across platforms and provided key insights into previously defined gene expression subtypes. The accompanied stand-alone tool will be freely available via our website. Future plans include extending the methodology presented to other data types and technological platforms. An interesting extension is also considering other pathologies to identify significant molecular heterogeneity.

Acknowledgements

Research carried out in the context of this study has been funded by the EU DICODE (Mastering Data-Intensive Collaboration and Decision Making) Collaborative Project (FP7, ICT- 2009.4.3, Contract No. 257184) and EU COST Action SeqAhead (BM1006).

References

Denison DGT, Adams NM, Holmes CC, Hand DJ (2002) Bayesian partition modelling, *Comput Stat Data Anal*, **38** (4): 475– 485.

Network TCGAR (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**: 1061-1068 doi: 10.1038/nature07385

Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**: 1807-1812. doi: 10.1126/science.1164382

Tyekucheva S, Marchionni L, Karchin R, and Parmigiani G (2011) Integrating diverse genomic data using gene sets. *Genome Biol*, **12**: R105. doi: 10.1186/gb-2011-12-10-r105

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C *et al.* (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**: i237-i245. doi: 10.1093/bioinformatics/btq182