Validation of Merging Techniques for Cancer Microarray Data Sets

Jonatan Taminau* Stijn Meganck Ann Nowé

Computational Modeling Lab
Vrije Universiteit Brussel
**jtaminau@vub.ac.be*

Abstract

There is a vast amount of gene expression data that has been gathered in microarray studies all over the world. Many of these studies use different experimentation plans, different platforms, different methodologies, etc. Merging information of different studies is an important part of current research in bioinformatics and several algorithms have been proposed recently. There is clearly a need to create large data sets which will allow more statistically relevant analysis in order to obtain more and better results in clinical and medical research. In this article we consistently describe several gene expression data merging techniques and apply them on several cancer microarray data sets.

1 Introduction

One of the main areas in bioinformatics is the analysis of microarray data, however in recent publications both the reproducibility and the validity of outcomes have been challenged [3]. Recently merging of data, i.e. combining different studies in order to increase the statistical power of the obtained results, and thus creating a larger and more reliable data set, has become a major point of research in bioinformatics.

2 Methods

In this article we study three test cases of increasing biological complexity.

- NCI60 is a collection of very well studied cell lines of nine different types of cancer. Both NCI60 studies used in this article use the same cell lines but were performed on different arrays or platforms and can therefore be seen as a relatively good benchmark case.
- For the Thyroid case we used two publicly available and two in-house thyroid cancer data sets. The thyroid cancer studies are more complicated since they use tissues instead of pure cells, thereby introduce more biological noise. There is however still a large and clear biological variance between tumors and normal tissues.
- We used a collection of eight breast cancer data sets as a last test case. Breast cancer data is more complicated than the other two cases since breast cancer is known as being a very heterogenous disease and there is a much smaller biological variation between the different breast cancer subtypes.

Every case consists of a number of studies, performed by different labs, using different technology and all studies within one case try to solve the same biological problem. For example, in the Thyroid case all studies have patients (samples) that are healthy or that have a tumor. In all studies the goal is to derive a prognostic model based on the gene expression data for the disease outcome. By combining the information from the different studies, one hopes to increase the statistical power of his model.

To combine or merge different studies several techniques are recently proposed in literature. In this article we will compare five such techniques in a consistent and extensive way in order to investigate if a certain merging technique performs consistently better than the other methods. The five techniques we consider can be very simple (RAW: do nothing, NORM: gene standardization, BMC: batch mean centering [5]) or more advanced and complex (DWD: Distance Weighted Discrimination [2], XPN: Cross Platform Normalization [4]).

3 Results and Conclusions

To validate the five different merging techniques we used a number of validation techniques and studied how consistent they are with each other. We started with a visual inspection of Multidimensional Scaling (MDS) plots of all samples after merging with each of the five merging methods. From these MDS plots we could easily observe that for the NCI60 and Thyroid case a high study-bias was present which could not be removed with the simple merging methods (RAW, NORM and BMC) and samples from the same study were clustered together, regardless of their biological annotation. For the more advanced methods (DWD and XPN) this was not the case however and samples belonging to the same biological class (e.g. Tumor vs Normal) were clustered together. For the Breast case this tendency was not significantly present due to two main reasons: (1) the biological annotation of the breast cancer samples (Estrogen Receptor Positive, ER+ vs Estrogen Receptor negative, ER-) is not so discriminative and (2) all breast cancer studies were performed on the same and stable platform, thereby introducing less study-bias than in the other cases.

Housekeeping genes are genes with relatively stable expression values between different tissues and conditions [1]. We used this property to again investigate the study-bias between two or more studies. The results we obtained after plotting all median values of the housekeeping genes before and after merging, were consistent with the conclusions we could derive from the MDS plots.

We also tried to assess the compatibility of the single studies after transformation caused by the five merging methods by performing cross study classification. We trained a SVM classifier on one transformed dataset (the largest) and then test it on the other transformed datasets. Remarkably, this validation indicated that simple normalization methods (NORM and BMC) performs beter than more complicated methods, which is in contradiction with the two previous validation outcomes.

The different validation techniques sometimes seem to contradict each other. This partly due to the fact that they test properties that are useful for different applications. A transformation that maximizes classification compatibility does not necessarily minimize the spread of a set of samples, etc. When working with microarray data it is therefore important to report why a specific validation technique was used. The different performance of the various merging methods on the presented cases opens the perspective of the development of a new method that can be applied generically, or of an intelligent algorithm to identify automatically which method is best for each case. We view this as part of our future work.

References

[1] Reija Autio et al. Comparison of affymetrix data normalization methods using 6,926 experiments across five array generations. *BMC Bioinformatics*, 10 Suppl 1:S24, Jan 2009.

[2] Monica Benito et al. Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–14, Jan 2004.

[3] Stefan Michiels et al. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–92, Jan 2005.

[4] Andrey A Shabalin et al. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, 24(9):1154–60, May 2008.

[5] Andrew H Sims et al. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. *BMC medical genomics*, 1:42, Jan 2008.