Analysis of CAMDA RNA-seq data with the knowlegde of protein domains in genes

Anna Leśniewska [1], Alicja Szabelska-Beręsewicz [2],
Joanna Zyprych-Walczak [2], Michał Okoniewski [3]

[1] Institute of Computer Science, Poznan University of Technology, [2] Department of Mathematical and Statistical Methods Poznan University of Life Sciences, [3] Scientific IT Services, ETH Zurich
May 19, 2017

Abstract

In RNA sequencing with short reads, it is often not possible to assign RNA fragment to a gene due to similarities in repeatable regions or protein domains. This may influence the downstream analysis. We have compiled the gene-domain database and used it for analysis to see the differences between the genes that share a domain versus the rest of the genes in the **Neuroblastoma dataset** . The major findings are:

- pairs of genes that share a domain have increased Pearson's correlation coefficients of counts
- the distribution of correlation coefficient for those pairs is leaning more towards the positive values for the for the smaller number of biological samples
- using diverse primary analysis counting strategies on non-CAMDA datasets suggests that the increased correlation reflects rather a real biological co-expression than sequence-based artifacts
- genes sharing a domain are expected to have a lower predictive power due to increased correlation, but with various type of classifiers the number of misclassified samples does not show yet an obvious dependence
- various classifiers perform in a very different way on the CAMDA data, which proves that clinical application of gene signatures from similar datasets may be difficult

We have to admit that outcomes are sometimes not following our intuition and experience from standard RNA-seq analysis. That is why we would like to present it at the CAMDA meeting and discuss there with the experts in the area.

Introduction

Many of the methods of data analysis in transcriptomics use specific measures for genes co-expression. One of the most obvious approaches is using Pearson's correlation coefficient. It is in fact the basis for popular heatmaps and hierarchical clustering of measured samples. However, as pointed out in the study [1] the positive correlations between the transcriptomic measurements may be an effect of real biological co-expression as well as artefactual correlation due to the technology specific issues. It is practically not possible to fully distinguish the increased correlation from both of the reasons. The study [1] has proven that in the Affymetrix techonology the increased correlation is seen for probesets that share probes with the same sequence.

In this analysis we propose an approach that is focused on gene structure and sequence composition in context of genome-wide analysis concerning the influence of protein domains, using the information from PFAM database [2]. The domains described in PFAM are the results of aminoacid-level analysis of sequences, thus not all the protein domain may have enough similarities on the nucleotide level of mRNA. Still, we use it as an initial approximation for sequence similarity, as creating a similar nucleotide database may be non-trivial, eg. the databse RFAM [3] includes only domains in non-coding sequences.

Data and Methods

Database of genes and domains. As the first step in the analysis the global table of functional domains and genes in which they are located was built from annotation databases. Appropriate database joins have been performed on the genomic coordinates of genes (AceView or Ensembl) and domains from Pfam. The data may be interpreted as a graph where the nodes are genes and domains. The graph consists of gene-domain-gene

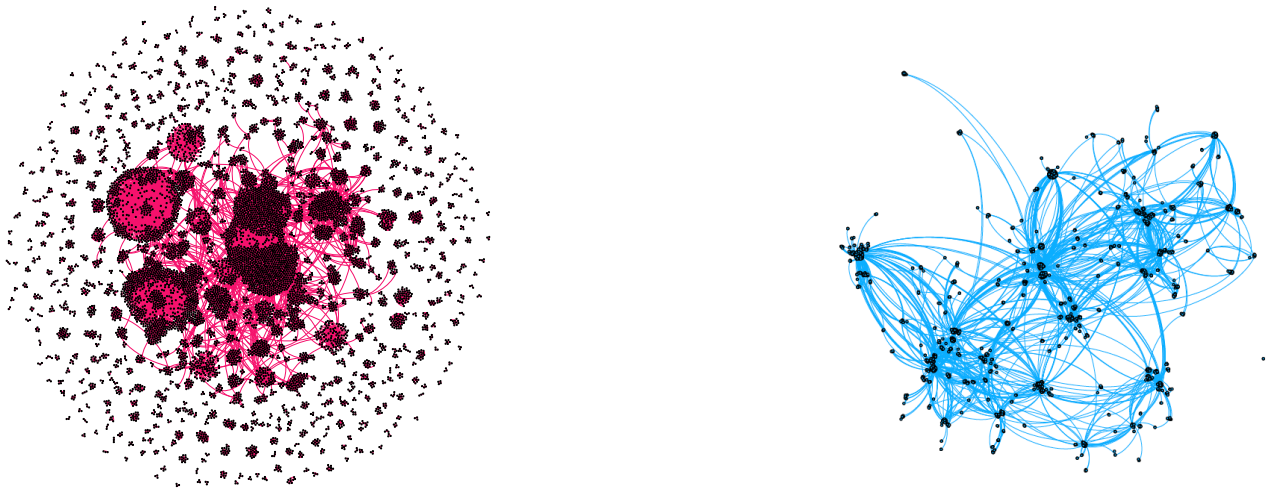

Figure 1: **Graphs visualized in Gephi, depicting genes interconnected with domains.** Left - the global picture, right - a single disconnected sub-graph. It shows that the interconnection of domains in the genes are not regular and trivial.

motifs, as a gene is connected with another gene always via a domain and vice versa. This builds the structural "galaxies" of gene families interconnected with domains (see the Figure 1).

From the CAMDA Neuroblastoma RNA-sequencing dataset we selected the subset of samples: 420 from Germany. We used in the analysis clinical genetic subgroups provided in CAMDA dataset as a grouping factor. Since some of the samples had unspecified subgroup we removed these samples. Dataset resulted with 302 samples. The genes were filtered to remove "almost empty" ones - those that include less than 2 samples with expression level above 5 normalized counts. After this step 18897 genes remained for further analysis.

The analysis included calculating co-expression coefficients for genes that share a structural domain. Expression values of a gene for different samples can be represented as a vector; thus calculating the co-expression measure between a pair of genes is the same as calculating the selected measure for two vectors of numbers. We checked most commonly used co-expression measures - Pearson's rank correlation coefficient, following the method from [1]. The distributions have been generated for the gene pairs sharing a domain and for "null" distribution of random gene pairs without a domain. Similar correlation coefficient distributions have been generated for Ensembl-based gene-domain-gene motifs with the numeric data from [4] which have been generated by counting with multiple mapping reads, with *-fraction* option and without counting multiple mapping reads using featureCounts from subread [5] package.

In addition machine learning approaches have been used for finding the importance of some differentially expressed genes. First, the differential expression was performed with edgeR approach [6]. Next, we have calculated the classification error taking into account differentially expressed genes with domains and without. The classification attribute was stage of the disease, with 4 values: ST4, ST4S, ST1, MNA. The input to the classifiers it was count data table of 50, 100 and 150 top differentially expressed genes with or without domains. We trained the classifiers based on these informations to find if the sample match to the particular clinical genetic subgroup. We used the following classifiers: k-nearest neighbor [7], support vector machine [8], the neural network [9] and random forest [10]. All of these classifiers are included in the MLInterfaces R package [11]. Leave One Out Cross Validation was used to calculate prediction errors counted as misclassification of samples. All the analyses have been carried out using R v3.2.0 and BioConductor v3.4.

Results

For the database integration done with Pfam and AceView, there are 20566 genes that share a domain, and 12666 genes without a domain. For analogous Ensembl joins there are 16923 genes with the domain and 41069 without.

We have calculated Pearson's correlation coefficient between the expression values of genes that share the same domain and between the expression values of genes that do not share the same domain. We plotted the histogram of correlation coefficients between the log value of counts for 25000 randomly chosen pairs of genes that share the same domains (green) or do not share any domain (red) (see Figure 2).

The increased correlation is visible. In the Affymetrix technology such phenomenon was explained partly by the artifacts of sequence similarity, partly by real biological co-expression [1]. In RNA-seq one can try to distinguish between those two types of effects on correlation by counting or not the multiple mapping reads, eg

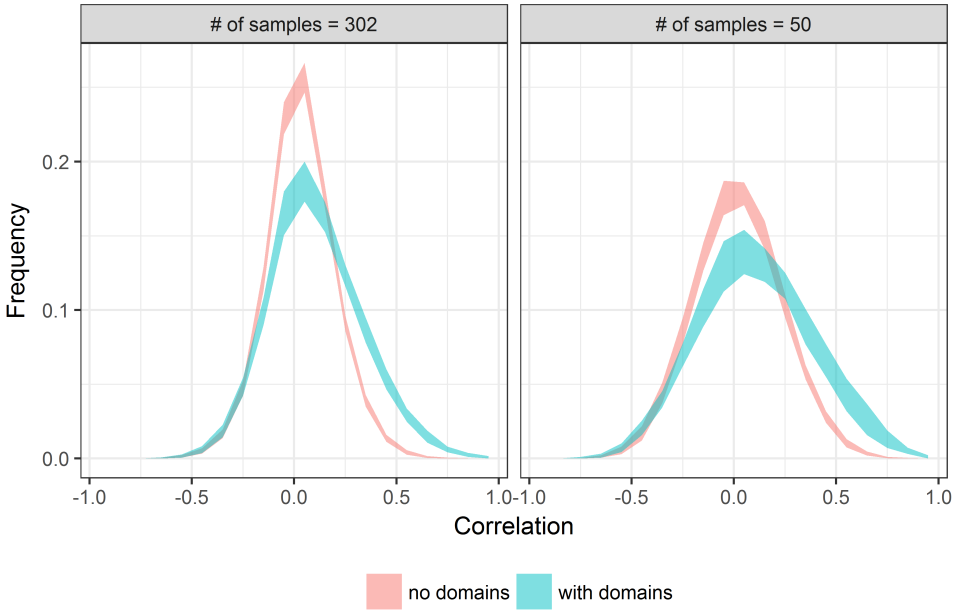

Figure 2: **An example of Pearson's correlation distribution for the pairs of genes with and without domains.** Red line is the histogram-based distribution for a random selection of gene pairs without domains. Green line is the distribution of the Pearson's correlation coefficient for the genes that share a PFAM domains. The increase in correlation is typically larger for smaller number of samples (right plot). Left - correlation calculated based on all samples; right - correlation calculated based on 50 randomly chosen samples

using featureCount [5]. Typically, the count tables of not multiple mapping genes include smaller numbers. In the CAMDA data we do not have access to the raw reads, thus the check with multiple mapped reads was done with another public dataset [4]. The not multi mapping counts have also a great deal of increased correlation (see Figure 3), so the origin of the effect is likely to be mostly purely biological, not artefactual. We have seen similar effects with several other RNA-seq datasets.

As a result of RNA-seq experiments, we obtain information on the expression of thousands of genes simultaneously. This explains the increase of the computational complexity involved in the classification process and has an adverse effect on the estimation of the prediction. In this part of our investigations we wanted to determine what is the prediction error in the case of classification. The gene selection process can help to obtain a subset of genes that can be used to distinguish different sample classes. Therefore, it is very important to carry out this step of analysis properly and as efficiently as possible.

The idea was to take into account the correlation structure of the genes in the selection process. We used the assumption stated in [12] that genes that are highly correlated with one to another belong to the same pathways or perform similar functions in cells. In the classification process we want to avoid the selection of highly correlated genes because they do not contribute much additional information to the classification [13] and also generate similar prediction errors in the process of discriminant analysis [14]. Therefore we used two subset of significant genes: with and without domains (see Figure 4).

In the last step, we would like to see which attributes (genes) were important. It is possible to do by building classification tree in randomForest method. Therefore, we have plotted variable importance plots based on randomForest package [10]. The first measure is computed from permuting OOB data. For each tree, the prediction error on the out-of-bag portion of the data is recorded (error rate for classification, MSE for regression). Then the same is done after permuting each predictor variable. The difference between the two are then averaged over all trees, and normalized by the standard deviation of the differences. If the standard deviation of the differences is equal to 0 for a variable, the division is not done (but the average is almost always equal to 0 in that case). The second measure is the total decrease in node impurities from splitting on the variable, averaged over all trees. For classification, the node impurity is measured by the Gini index. For regression, it is measured by residual sum of squares (see Figure 5).

Funding This research has been partly supported by Polish National Science Center grants: 2015/17/D/ST6/04063 and 2014/13/B/NZ2/01248.

References

[1] Michał J Okoniewski and Crispin J Miller. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC bioinformatics*, 7(1):276, 2006.

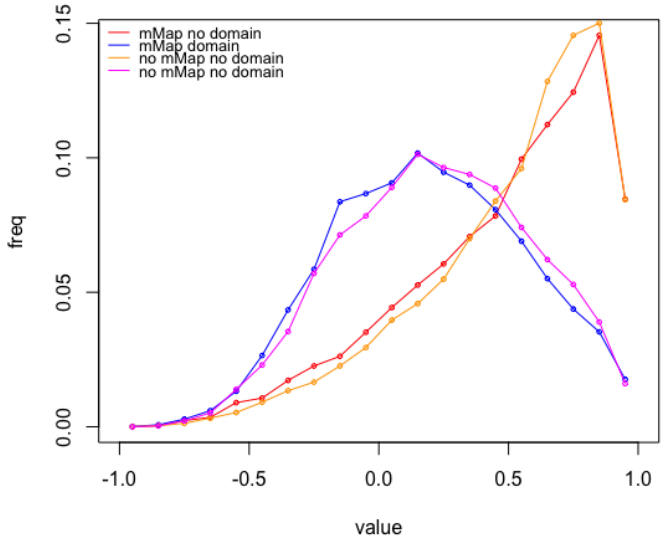

Figure 3: **The histogram-based distributions(with 20 buckets) of Pearson's correlation coefficient for random and domain-connected pairs of genes in the [4] dataset.** The counting takes into account reads with multiple mapping and only specific reads. The multi-map removed case has a smaller difference between domain and non-domain pairs, but the difference is small. This suggests that the origin of increased correlation is real biological co-expression.

[2] Robert D. Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L. L. Sonnhammer, John Tate, and Marco Punta. Pfam: the protein families database. *Nucleic Acids Research*, 42(D1):D222, 2014.

[3] Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna, and Sean R. Eddy. Rfam: an rna family database. *Nucleic Acids Research*, 31(1):439, 2003.

[4] Yuwen Liu, Jie Zhou, and Kevin P. White. Rna-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301, 2014.

[5] Yang Liao, Gordon K. Smyth, and Wei Shi. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):e108, 2013.

[6] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[7] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[8] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017. R package version 1.6-8.

[9] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[10] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[11] VJ Carey, R Gentleman, J Mar, Vertrees cfJ, and Gatto L. Mlinterfaces: Uniform interfaces to r machine learning procedures for data in bioconductor containers, October 2016.

[12] Mayer Alvo, Zhongzhu Liu, Andrew Williams, and Carole Yauk. Testing for mean and correlation changes in microarray experiments: an application for pathway analysis. *BMC Bioinformatics*, 11, 2010. http://www.biomedcentral.com/1471-2105/11/60.

[13] Dechang Chen, Zhenqiu Liu, Xiaobin Ma, and Dong Hua. Selecting genes by test statistics. *J Biomed Biotechnol.*, 2:132–138, 2005. doi: 10.1155/JBB.2005.132.

[14] J. Jaeger, R. Sengupta, and W.L. Ruzzo. Improved gene selection for classification of microarrays. *Pacific Symposium on Biocomputing*, 8:53–64, 2003.

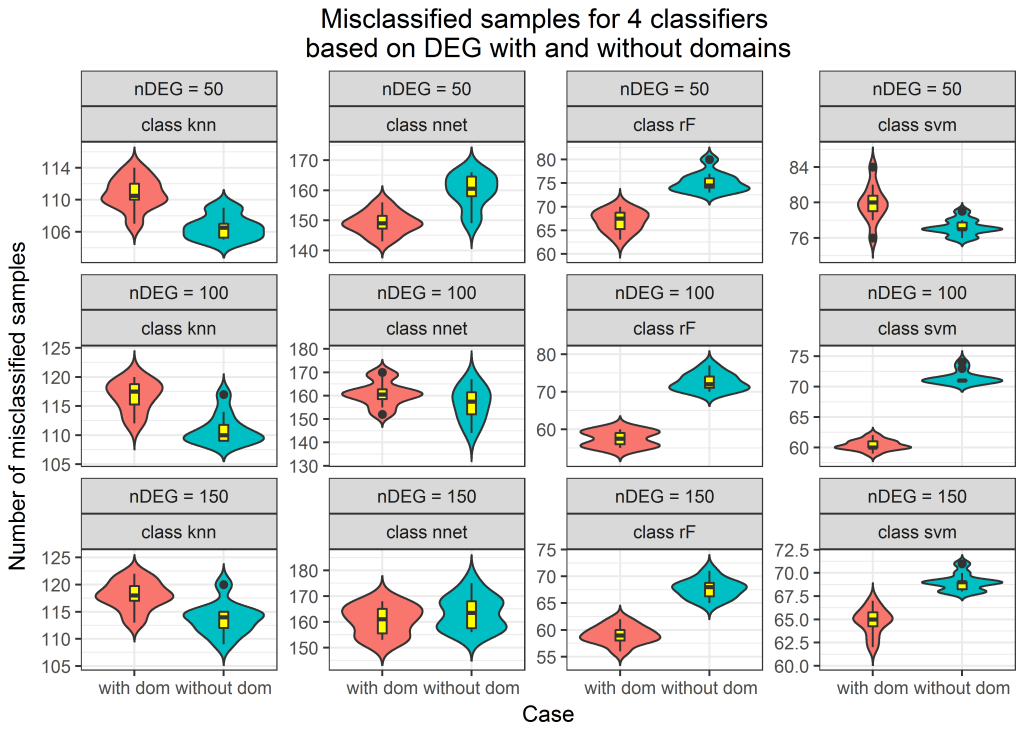

Figure 4: **Violinplot of misclassified samples for 4 classifiers based on DEG with and without domains.** Total number of samples was 302. There are differences between the domain and non-domain genes, but they differ depending on the type of classifier. More tests are needed in that area - on the selection of domain associated genes as the input to classifiers and its effects.

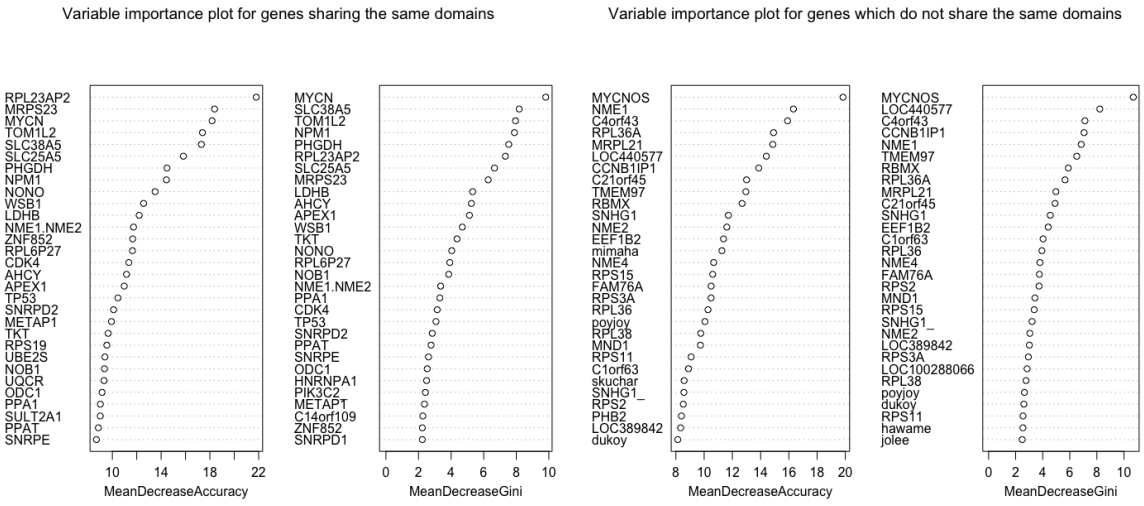

Figure 5: **Variable importance plot of the attributes for 30 top genes with and without domains - result of top decisive attributes in the random forest classifier**