TEMP069

BENCHMARKING PRE-PROCESSING AND BATCH EFFECT REMOVAL METHODS FOR INSILICO DB: GENOMICS BIG DATA INFRASTRUCTURE

Quentin De Clerck (1), Ann Nowe (1), Alain Coletta (2), Dipankar Sengupta (1)

1) Como, Vrije Universiteit Brussel, Pleinlaan, 1050 Brussels, Belgium
2) InSilico Genomics S.A., Avenue Adolphe Buyl, 87, B-1050 Brussels, Belgium

Genomics hold great promises for the future of medicine, because of its potential in discovering the genetic origin of diseases. Usually, genomic data are generated from varied technologies like microarray, RNA sequencing (RNA-Seq), etc., that uses different experimental settings. The principal challenge in analysing such diversified data is, people having biomedical expertise often do not have an apt access, making it impossible for them to interpret the results. Thus, the need of the hour is to enable researchers, to access and interpret the information enclosed in the ensemble of genomes by providing them with tools that can be easily manipulated. One of such tools is InSilico DB[1,2]. It comprises of a web-based genomic dataset hub, that enables users to manage genomic data in a user-friendly way[2]. In this work, we are trying to enrich the InSilico DB platform by adding enhanced pre-processing methods accessible from the InSilicoDb R-package[1].

In this study, we pipeline two recent pre-processing techniques, beside frozen Robust Multiarray Analysis (fRMA)[3], that solve for the necessity of a reference dataset. The first method, Single Channel Array Normalization (SCAN)[4] uses intrinsic information of the sample for estimating the background noise and normalizing the microarray data. Since, the SCAN method only uses information from the sample itself, it allows to pre-process all the Affymetrix data sets from InSilico DB. The second method, Universal exPression Code (UPC)[5], supports a larger variety of data types making it a good option for combining datasets originating from different technologies like microarray and RNA-Seq. Further, we propose to modify the genomic pipeline to generate UPC values for the RNA-Seq data type. RNA-Seq data files are heavy, therefore, can be categorized as Big Data, and aligning them to a reference genome is time consuming. Therefore, the Apache Hadoop[6] framework is being used to manage the RNA-Seq data and its MapReduce implementation for alignment to a reference genome.

The proposed pipeline along with the pre-processing techniques, were tested in combination of batch effect removal methods (NONE, COMBAT, BMC, GENENORM)[7]. They were applied to a dataset[7], in which microarray and RNA-Seq datasets were grouped together that are suitable for merging. Principal Component Analysis (PCA) was used for merging of data. For benchmarking, we used the transformed gene expressions of a batch effect removal method as training data for decision trees. Our hypothesis for finding the best combination of batch effect removal and merging approach was the following: If the data is successfully merged, retrieving the study through classification should be hard, while classifying healthy versus disease should become easier. For each combination 20 trees were generated using a random selection of the available data: 10 each for the study and disease The complexity of the generated tree is used as a measure, which is determined by multiplying the total number of nodes and the depth of the tree. The choice of the stated criterion is motivated by the fact that complex tree structures have more nodes, whereas depth is considered for the flatness of the tree. A deeper tree indicates a higher difficulty for the classification of the data. COMBAT scored the best of all batch effect removal methods when combined with SCAN and fRMA, having high and low complexity respectively, to retrieve study and disease trees. This study enhances the feasibility to analyse, both legacy microarray data with newer RNA-Seq data together and find relevant biological information.

References

1) Taminau, J., et al. InSilicoDb: an R/Bioconductor package for accessing human Affymetrix expert-curated datasets from GEO. Bioinformatics, 27(22):3204-5 (2011).
2) Coletta, A., et al. InSilico DB genomic datasets hub: an efficient starting point for analyzing genome-wide studies in GenePattern, Integrative Genomics Viewer, and R/Bioconductor. Genome Biology, 13(11):R104 (2012).
3) McCall, M.N., Bolstad, B.M., and Irizarry, R.A. Frozen robust multiarray analysis (fRMA). Biostatistics, 11(2):242–253(2010).
4) Piccolo, S.R., et al. A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics, 100(6):337–344 (2012).
5) Piccolo, S.R., et al. Multiplatform single-sample estimates of transcriptional activation. Proceedings of the National Academy of Sciences, 110(44):17778–17783 (2013).
6) White, T. Hadoop: The Definitive Guide. O'Reilly Media, 1st edition (2009).
7) Taminau, J., et al. Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages. BMC Bioinformatics,13(1):335 (2012).