

ZB MED - Informationszentrum Lebenswissenschaften

14th Day of the Doctoral Student

Seyfert, Hans-Martin

Dummerstorf, 2013

[Preface]

urn:nbn:de:hbz:38m:1-30065

“If you can’t explain it to a six year old, you don’t understand it yourself.”

(Albert Einstein)

The Leibniz Institute for Farm Animal Biology is attaching special importance to the promotion of the new generation of academics. The ‘Day of Doctoral Student’, which we are holding annually since 2000, opens our PhD students the opportunity to report the other doctoral students, scientists and guests their achievements and progress in their scientific work.

‘Promotion by demanding’ is the motto of the FBN in supervising the PhD students. In this regard, the ‘Day of Doctoral Student’ has meanwhile proved as a valuable exhibition-competition of our young academics. Within two groups (‘Beginners’, and ‘Advanced students’) selected PhD projects are presented in a short form. The presentations are evaluated by a jury consisting of the directors of the Institutes and lecturers present and the best presentation of each group will be awarded.

This slightly formalised presentation gives the up-and-coming academics the opportunity to practise an exactly terminated structured scientific talk. These presentations serve to preselect FBN-contributions for participation in important scientific events, such as the annual meetings of the Societies for Animal Breeding (DGfZ/GfT), Reproduction medicine (DVG/DGRM) and Nutrition Physiology (GfE).

Apart from finishing the doctoral thesis in time, contemporary publication of their results is required from the young scientists. With the current issue of the journal ‘Schriftenreihe des FBN’ we give the advanced students the opportunity to publish their results in a citable journal and acquire first experiences in publication of scientific results.

As researchers please keep in mind that you never fail until you stop trying. I wish success to all PhD students in their further career as well as personal luck and well-being.

Yours

A handwritten signature in cursive script, reading "Manfred Schwerin".

Prof. Dr. Manfred Schwerin

Director

A workflow to investigate the additional usage of protein-protein-interaction data for the classification of oestrogen receptor status in breast cancer based on gene expression

Clemens Harder^{1,2}, Georg Füllen² and Dirk Repsilber¹

¹Leibniz Institute for Farm Animal Biology (FBN), Dummerstorf, ²Institute for Biostatistics and Informatics in Medicine and Ageing Research (IBIMA), University of Rostock, Rostock

1 Introduction

Gene expression profiles are used for prediction/classification of phenotypes. Each gene represented in the expression profiles used for classification is denoted as feature. To increase the prediction accuracy researchers produce artificial features, for example profiles calculated from gene pairs [1, 2]. These newly constructed features can lead to better classification results than features based on single gene expression profiles. In the last years researchers started to combine primary data, such as gene expression profiles, and secondary knowledge, like Protein-Protein-Interaction (PPI) data or biological pathways, to gain more insights into the underlying biological processes [3]. There exist both, promising articles [4, 5] as well as critical ones concerning the additional usage of interaction data for phenotype classification [6]. The following work combines gene expression data and secondary knowledge in the form of interacting gene pairs extracted from the String Protein-Protein-Interaction data base [7]. We investigated the impact of using these biological interactions on the prediction accuracy of the classification of the oestrogen receptor (ER) status in breast cancer applying two standard classifiers. The resulting workflow is available for other OMICs classification problems and as a primer for further methodological research.

2 Materials and methods

Except for the data acquisition all of the following steps were implemented using R [8].

Data Sets: Seven data sets were compiled, 6 from a Gene Expression Omnibus [9] search for breast cancer (GSE7390 [10], GSE6532 [11], GSE4922 [12], GSE3494 [13], GSE2990 [14], GSE25066 [15]) and one from Hess et al. [16]. Each of these data sets consist of more than 33 samples for each class (ER positive, ER negative) and were generated on the Affymetrix platform GPL96.

Raw Data Processing: The quality of the raw expression data were checked with SimpleAffy from Bioconductor [17]. Arrays not fulfilling the metrics used by Popovici [18] were excluded. The RMA algorithm [19] was applied for within array normalization and the probe sets were mapped with a custom brainArray [20] cdf-file („hgu133ahsentrezgcd“ version 16.0.0) to Entrez Gene identifiers [21]. The expression matrices were normalized for between-array effects by a quantile normalization.

Balancing Class Sizes: If one class had more samples, we randomly reduced its sample number, so that both classes had equal number of samples.