BIOINFORMATICS

Vol. 1 no. 1 2015
Pages 1–10

Finding relevant features for Breast invasive carcinoma (BRCA) with classification methods (multiple linear regression and support vector machine) and network based methods (Pearson correlation and partial correlation).

Irene Ziska [1] and Florian Heyl [2]

[1]Freie Universität Berlin, Berlin Germany
[2]Freie Universität Berlin, Berlin Germany

ABSTRACT

There exists a variety of clinical data which can be used to identify features related to specific types of cancer. To improve the results of a therapy the stage of the cancer has to be known. The patient could get a treatment especially designed for this type and stage of cancer. The problem that must be faced is the large amount of data which grows exponentially every year. In this project we introduce a pipeline which is doing these two steps. First of all, it classifies what stage of cancer the patient has and afterwards, it checks if specific genes for this type of cancer are activated or inhibited. We started to evaluated data for breast invasive carcinoma (BRCA) from the Cancer Genome Atlas (National Cancer Institute , n.d.). We found two parameters - the number of lymph nodes positive by hematoxylin and eosin staining and light microscopy (H&E staining and lm) and the her2/cent17-ratio - which can be used to classify the patients into two cancer types, spreading and non-spreading. We also found three gene expression networks based on Pearson and partial correlation which could play an important role for a BRCA. For the functional annotation we used DAVID (David , n.d.) and EnrichtNET (EnrichNet , n.d.). Unfortunately the significance values for theses genes were really bad. We additionally analyzed proteins related to BRCA and found again three networks based on Pearson and partial correlation. For that analysis we used Gene Ontology (GO , n.d.) for functional annotations.

Contact:
i.ziska@fu-berlin.de
florian.heyl@fu-berlin.de

1 INTRODUCTION

1.1 Types of breast cancer

Breast cancer is one of the most frequently diagnosed cancer types in women. There are 458.000 deaths per year worldwide making it the most common cause of female cancer death (Eccles *et al*., 2013). Invasive or infiltrating ductal carcinoma (IDC) is the most common type (American Cancer Society , n.d.). Another one is Ductal carcinoma in situ (DCIS) which is also known as intraductal carcinoma. This type is considered non-invasive or pre-invasive breast cancer. The difference between DCIS and invasive cancer is that the cells have not spread through the walls of the ducts into the surrounding breast tissue (American Cancer Society , n.d.). It is possible that a DCIS evolves and become an invasive cancer. Still, no classifier could be developed that predicts the outcome or the chance that a certain case evolves to an invasive or non-invasive cancer.

There exist also other sub-types of breast cancer. There is no general therapy as a cure. Hence, each sub-type needs an individual treatment. That is why there are two main types of staging systems for cancer. They should help to decide which treatment is the best for the given patient. The first one is the TNM system. The second one which plays a major role in the analysis of this project is the number system (Cancer Research UK , n.d.). The number system is a language that describes the size and the level of spread of cancers. Currently, the problem is that the systems are not strictly standardized. Therefore, different hospitals and different blood cancers or lymph system cancers have their own staging systems. Most types of cancer have four stages (Cancer Research UK , n.d.).

- Stage 1: The cancer is relatively small and not spreading. It is contained within the organ where it first occurred.
- Stage 2: Same as Stage 1 but the tumor is larger. For some cancer types it could additionally mean that the carcinoma spread into lymph nodes close to the tumor.
- Stage 3: The tumor is larger than in stage 2. The carcinoma may have started to spread into surrounding tissues and the lymph nodes in the area contain cancer cells.
- Stage 4: The cancer is even larger than in stage 3 and has spread to another body part from where it started. This cancer can be classified as metastatic.

Sometime there are some letters next to the stage number (A, B or C). These letters divide the stages again and allow a more detailed classification.

1.2 Features associated with breast cancer

Hematoxylin and eosin (H&E) stains are very important for the recognition of various tissue types and morphologic changes which is part of cancer diagnosis. Hematoxylin has a deep blue-purple color and stains nucleic acids by a complex (Fischer *et al*., 2008).

Eosin is pink and stains proteins non-specifically (Fischer *et al*., 2008). With these two dyes it is possible to find diagnostically very important cell-type- and cancer-type-specific patterns. With a higher cancer stage the number of positive (abnormal or malign) lymph nodes should increases due to the fact that a cancer has a higher chance to spread.

HER2 overexpression is a common feature of DCIS (Curigliano *et al*., 2015). The overexpression, respectively the amplification of HER2 goes up to 60 % to 70 % in patients with high-grade DCIS. This results in an occurrence of necrosis and p53 mutation. It is also associated with the expression of hormone receptors. That is why HER2 could be a valid therapeutic target in invasive breast cancer and could be additionally beneficial for patients with DCIS (Curigliano *et al*., 2015). The HER2 status can be determined by FISH (fluorescent in *situ* hybridization) or by calculating the **her2/cent17 (or her2/cen17) ratio** (Chibon *et al*., 2009). The ratio is calculated by dividing the number of HER2 copies by the number of chromosomes 17. A polysomic status (having an extra copy of one or more chromosomes) of chromosome 17 are known to be associated with pathological and clinical factors in breast cancer. Therefore the ration could help to predict the cancer stage of individual patients.

2 APPROACH

The goal of this project was to built a pipeline which classifies features (i.e. parameters, genes, proteins) that are specifically associated with a breast invasive carcinoma (BRCA). For that purpose a pipeline was created that can be split into two major parts (see figure 1). The first part constructs a classifier based on a regression model and a support vector machine (SVM). The second part evaluates genes that could play a role in this specific type of cancer by considering gene expression data. The same is done for protein expression data. The data for both parts has been derived from the Cancer Genome Atlas (National Cancer Institute , n.d.). A detailed description of the data can be found in the material section.

2.1 Multiple linear regression and SVM

In the first part of the pipeline clinical data of patients from the Cancer Genome Atlas (National Cancer Institute , n.d.), which where part in the BRCA study, was considered. This set contained 1092 patients with data for several clinical parameters like the cancer stage of the patient. The goal was to create a classifier based on clinical parameters to predict the cancer stage of a patient. Therefore, we searched for parameters which are obviously in some kind of functional relation to the cancer stage of a patient. With that approach we found three parameters: number of lymph nodes positive by H&E staining and lm, her2/cent17 ratio and the days until last follow up. Afterwards, we generalized the cancer stages. We created two groups of patients, a stage 1 cancer (non-spreading) type and a stage 2 cancer (spreading) type. For stage 1 we included all patients having the cancer stages I, IA, IB or IC. Stage 2 contained patients having the cancer stages II, III or IV (with sub-types A,B and C). The multiple linear regression and the SVM we used can be found in the methods. To improve the power of the prediction a k-fold cross-validation (see methods) was done. As result 10 models were built. The model with the smallest prediction error was picked. This model was the final classifier.

2.2 Gene and protein expression networks

In the second part we tried to find genes that are important for a BRCA. In order to find these genes, gene expression data of 135 patients was processed. The data contained log2 ratio for 17814 genes. First of all, the Pearson correlation between these genes was computed. With the help of the correlation values an undirected network was constructed. In this network an edge between two genes means that these are connected, i.e that there is a dependency between these two genes. As threshold for an edge we used a correlation value of 0.9, respectively -0.9. For all edges the weight was set to 1 to imitate an unweighted network. In the constructed network we detected communities. The three largest communities were saved for further investigation. In the last step the data of one specific patient was joined with one of these modules at a time. As result three tables were created. These tables consists of the genes of one module and for each gene the mean ratio across all patients which we used for the training. Additionally, the specific ratio for the given patient can be found. There are some downsides to Pearson correlation. That is why we repeated this approach with a network that was constructed based on partial correlation. Afterards, the results of both versions were compared.

In addition to the gene expression networks we also built a protein expression network. For that purpose we used protein expression data of 20 healthy and 20 cancer patients. There was one problem with the given data because the files contained different proteins. Therefore, we included only proteins which were present in all files (148 proteins remained). We set the correlation threshold for the protein network to 0.75, respectively -0.75. The steps were the same as for the gene expression networks. Again both, Pearson correlation and partial correlation, were used to construct a network.

3 METHODS

3.1 Multiple Linear Regression

Given a set of input data (x,y), multiple linear regression tries to find a linear function (equation 1) which best fits the provided input data by localizing a vector w such that the sum of the squared residuals (equation 2) is minimized (FlinkML , n.d.).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_p x_{ip} + \varepsilon_i \quad \text{for} \quad i = 1, 2, ...n. \tag{1}$$

$$S(w) = \sum_{i=1} (y - w^T x_i)^2 \tag{2}$$

One can obtain the formulation in matrix notation (equation 3) where the equation 4 is a closed form solution.

$$w^* = \arg\min_w (y - Xw)^2 \tag{3}$$

$$w^* = (X^T X)^{-1} X^T y \tag{4}$$

The gradient for a given point x_i is given by equation 5 with the scaling $y = \frac{s}{\sqrt{j}}$ where s is the initial step size and j being the current iteration number (FlinkML , n.d.).

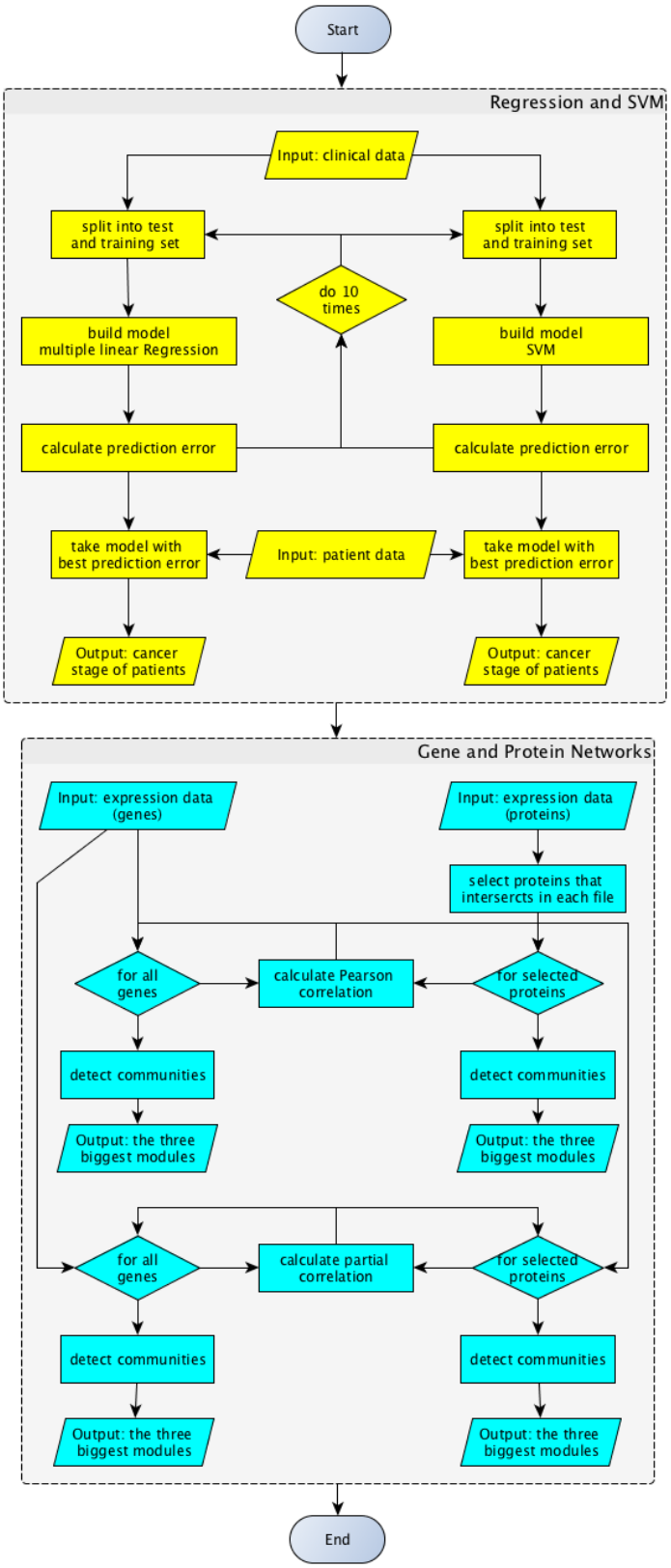

Fig. 1. The flowchart displays the workflow of our pipeline.

$$\bigtriangledown S(w, x_i) = 2(w^T x_i - y)x_i \tag{5}$$

The initial step size for the gradient descent method controls how far the gradient descent method moves in the opposite direction of the gradient (FlinkML , n.d.). We set the value to 0.001 for our pipeline. The multiple linear regression algorithm computes either a fixed number of gradient iterations or terminates based on a dynamic convergence criterion (FlinkML , n.d.). The convergence criterion is the relative change in the sum of squared residuals (equation 6)

$$\frac{S_{k-1} - S_k}{S_{k-1}} < p \tag{6}$$

We set convergence threshold to 0.001 and the number of iteration of the gradient to 1.000.000.000.

3.2 Support Vector Machine

The pipeline includes an SVM with soft-margin using the communication-efficient distributed dual coordinate ascent algorithm with hinge-loss function solving the minimization problem (see equation 7, FlinkML (n.d.)).

$$\min_{w \in \mathbb{R}^d} \frac{\lambda}{2}||w||^2 + \frac{1}{n}\sum_{i=1}^{n} l_i(w^T x_i) \tag{7}$$

The weight vector w, the regularization constant λ, the data points $x_i \in \mathbb{R}^d$ and l_i being the convex loss functions, which can also depend on the labels $y_i \in \mathbb{R}$, are part of the minimization problem (FlinkML , n.d.). In the current implementation the regularizer is the $\ell 2$-norm and the loss functions are the hinge-loss functions (see equation 8 FlinkML (n.d.)). We set the regularization constant to 0.002 and the initial step size to 0.1.

$$l_i = \max(0, 1 - y_i w^T x_i) \tag{8}$$

The minimization problem is solved by applying stochastic dual coordinate ascent (SDCA) (FlinkML , n.d.). The CoCoA algorithm calculates several iterations of SDCA locally on a data block before merging the local updates into a valid global state which is then redistributed to the different data partitions where the next round of local SDCA iterations is then executed (FlinkML , n.d.). We set the number of blocks into which the input data will be split to 10. The number of outer iterations and local SDCA iterations control the overall network costs, i.e. the outer iterations defines how often the SDCA method is applied to the blocked data (FlinkML , n.d.). We set the maximum number of iterations to 100. The local SDCA iterations define how many data points are drawn from each local data block to calculate the stochastic dual coordinate ascent (FlinkML , n.d.). We set the maximum number of SDCA iterations to 100.

3.3 k-fold Crossvalidation

For the purpose of improving the predictions we developed a 10-fold cross-validation for both the multiple linear regression and the SVM. In this step we defined a loop where in every loop the data is split up in a test and training set. We prevented a preference of patients with a specific cancer stage by randomly shuffling the list of patients. In each loop 10 % (k) of the shuffled list is used for the validation, the rest of the patients (1-k) is used for training. For each model the relative prediction error (RPE) is calculated (see equation 9) such that we obtain ten models with ten different prediction errors.

$$RPE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N}} \tag{9}$$

3.4 Model picking

After the k-fold cross-validation we picked the model which had the lowest prediction errors. This model was used to make the predictions. We looked for the patients we were interested in and answered the question which cancer stage they have.

3.5 Pearson correlation

The Pearson correlation describes the dependency between two variables. It can only measure linear dependencies between variables. The correlation value can be between -1 and +1. A correlation value of zero means that there is no correlation between the two variables, a correlation value of +1 describes a total positive correlation and -1 a total negative correlation. A positive correlation means that increasing values in the first variable correlates with increasing values of the second variable. A negative correlation means that increasing values in the first variable correlates with decreasing values of the second variable. The correlation coefficient is computed by dividing the covariance of the two variables by the product of the standard deviations:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{10}$$

The main problem of the Pearson correlation is that it cannot find non-linear dependencies. Additionally, if one gene regulates two other genes, Pearson correlation will compute a high dependency between the two regulated genes although there are independent of each other and only connected by their common regulator. Therefore, Pearson correlation does not take into account the effect of other variables when computing the correlation coefficient.

3.6 Partial correlation

In contrast to Pearson correlation, partial correlation removes the effect of other variables when computing the correlation coefficient. It is possible to compute the partial correlation between two variables given one single variable, given a subset of variables or given the rest of all variables. In this project the partial correlation between two variables given one single variables was used. It can be derived from the correlation matrix. The partial correlation between X and Y given Z is:

$$r_{xy|z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} \tag{11}$$

By concentrating on the significant correlations Ψ (within the defined threshold of 0.9 or -0.9), we applied to each of these correlation the calculation of the partial correlation. For each correlation α (X and Y) with $\alpha \in \Psi$ we calculated the partial

correlation with each gene or protein Z of the rest of the genes or proteins. Therefore, by calculating each partial correlation for α with the rest Z we created a set of partial correlations Ω. Then, we looked if one of the partial correlations in Ω is lower than a second threshold of 0.1. In theory if there would exist a partial correlation for α in Ω which is lower than 0.1, then one can infer that the correlation of X and Y can be explained by a third variable. This would then indicate that X and Y are not truly correlated and that the correlation can be removed from Ψ. So we removed all correlations which did not fulfill our conditions.

3.7 Community detection

To be able to find subsets of connected nodes in a network community detection can be used. We used the Community Detection example of Flink (Flink CommDecEx , n.d.). The number of iterations was set to 1000.

3.8 Table API

Flink's Table API can be used to read in and process structured text files. It allows to specify operations that are similar to SQL expressions (TableAPI , n.d.). In this project the Table API was used multiple times to easily access the data and to join results.

4 MATERIALS

We have taken three major data sets of BRCA from the Cancer Genome Atlas (National Cancer Institute , n.d.). There exist 1098 patients with BRCA. The patients can be divided in TN and NT. TN patients have a BRCA and match to the data of a person without the tumor. NT patients do not have a BRCA but they match to the data of a person with the tumor.

4.1 Linear Regression and SVM

For multiple linear regression and the SVM we looked at the clinical data of the whole patient pool (1098 patients). The data was produced by the Nationwide Children's Hospital (nationwidechildrens.org) Biospecimen Core Resource (BCR) using the Biospecimen Metadata - Complete Set platform (National Cancer Institute , n.d.). Three parameters to predict the cancer stage of the patients were investigated (see approach). Some data was not available for some patients. That is why the data set was reduced to 205 patients.

4.2 Gene expression network

To compute the correlation between genes we used the expression level 3 data of batch 8.0.0. The batch consists of 135 patients. This data contains log ratios for 17814 different genes. For each gene a log2 ratio is given. The ratio is computed between the tumor sample and a normal reference. For a healthy patient the ratio is 1 and therefore the log2 ratio is zero. A log2 ratio that is smaller than zero corresponds to a decreased gene expression. A log2 ratio that is greater than zero corresponds to an increase in gene expression.

4.3 Protein expression network

In TGCA also data for protein expression can be found. Protein expression is only measured for a selected set of proteins. Lots of them are known to be important in the development of cancer and in tumor suppression. The data was produced by MD Anderson Center (mdanderson.org). There are three levels for protein expression in TGCA. In this project we used the level 3 data. In level 3 the data has been normalized. It can be derived from level 2 data by subtracting the median for each protein across all the samples from values within each protein and by afterwards subtracting the median for each sample across all proteins from values within each sample (TCPA , n.d.). In this project level 3 data of 20 NT patients and 20 TN patients from different batches was used to compute a correlation between proteins based on their expression level.

5 RESULTS

5.1 Multiple linear regression and SVM

The data for the multiple linear regression and SVM (see materials) where prepared beforehand such that only necessary data (columns) where extracted and saved in a separated file. The pipeline reads in the data and split it into a training and test set for both the multiple linear regression and the SVM (see methods). The data for the three parameters (number of lymph nodes positive by H&E staining and lm, her2/cent17 ratio and number of days until last follow up) can be seen in the figure 2 A-C. One can clearly see that the number of lymph nodes positive by H&E staining and lm and the her2/cent17 ratio increases in a type 2 (spreading) cancer. The number of days until last follow up on the other hand is not so clearly correlated. We have tried to built up a model with all three parameters but we had to remove the parameter 'number of days until last follow up'. With the third parameter the models were to bad. So we build ten models with the first and second parameter via a k-fold cross validation (see methods) for the multiple regression and the SVM and picked the model with the best prediction error. The model for the multiple regression had a relative prediction error of about 1.533. The model for the SVM had a relative prediction error of about 0.775. The model of the multiple linear regression has the form: $y = x1 \cdot 0.070 + x2 \cdot 0.057 + 0.035$. The parameter 'number of lymph nodes positive by H&E staining and lm' has a weight of 0.070 and the parameter 'her2/cent17-ratio' has a weight of 0.057. We saved in a separated file the predictions based on the testing data for the best model (see table 1). With these models (for multiple linear regression and SVM) we made our predictions for patients of another data set (see table 2) regarding the cancer stage.

5.2 Gene expression networks

For both approaches, Pearson and partial correlation, different results could be obtained. Therefore, the results are presented separately.

Pearson correlation

After computing the Pearson correlation for all 17.814 genes, 3824 correlations could be found with a correlation coefficient greater than 0.9 or less than -0.9. Hence, the resulting network contained 3824 edges. The largest community that could be found consisted

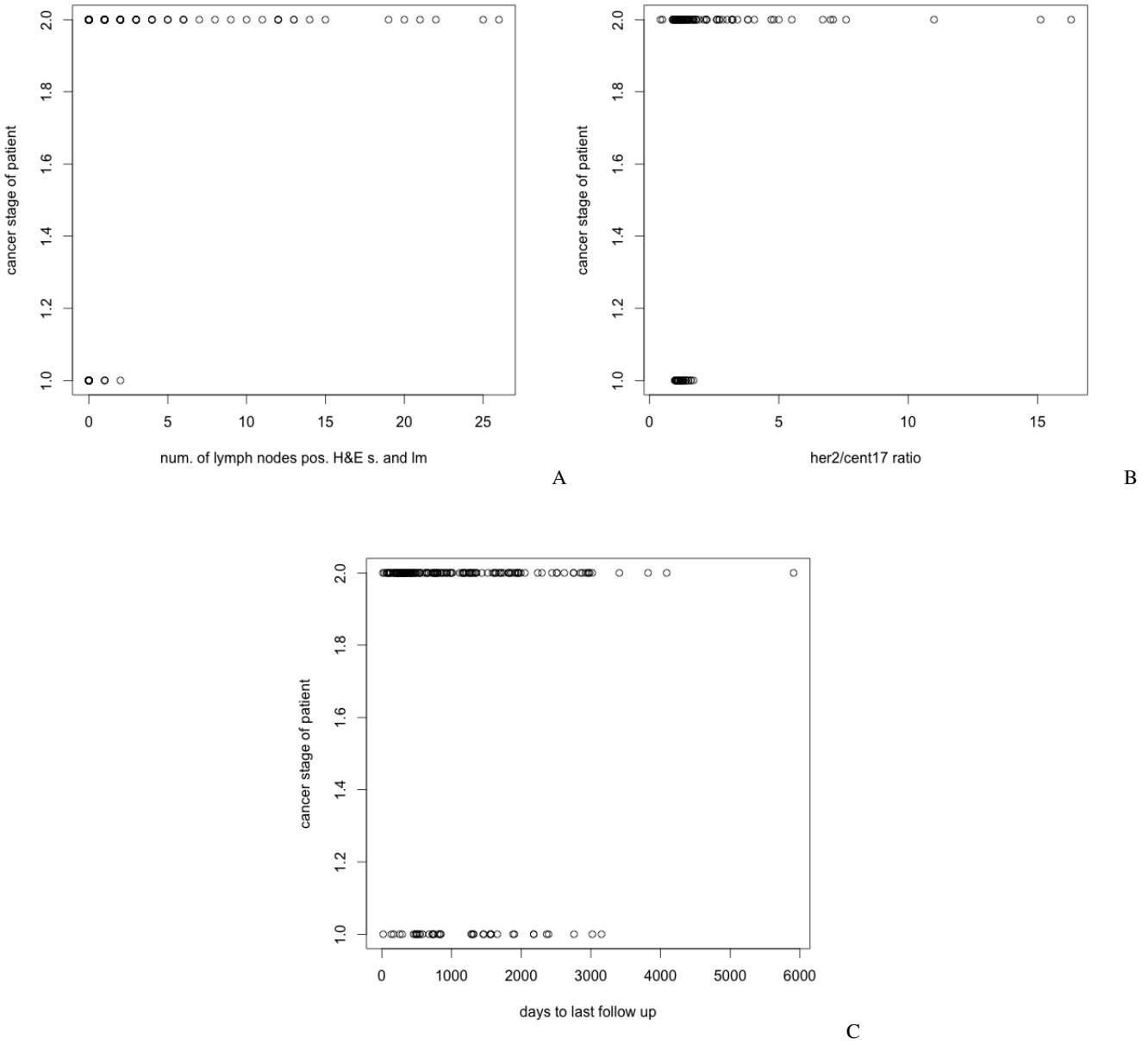

Fig. 2. Diagram A displays the number of lymph nodes positive by H&E staining and lm associated with the cancer stage 1 and 2 of the patients. There is obiously a tendence of an increasing number of lymphe nodes with a type 2 (spreading) cancer. Diagram B displays the her2/cent17 ratio associated with the cancer stage 1 and 2 of the patients. The ratio is often greater if the patient has a type 2 (spreading) cancer. The diagram displays the days of the last follow up (last contact to the patient) associated with the cancer stage 1 and 2 of the patients. There is not a clear tendence for the parameter.

of 101 genes, the second largest of 74 and the third largest of 56. In order to analyze the different modules for each module the gene symbols were put into DAVID (David , n.d.) and into EnrichNet (EnrichNet , n.d.). In DAVID the functional annotation clustering was used. The classification stringency was set to medium. The first three clusters were considered.

For the first module there could be found and overlap of 10 genes that are related to the regulation of the transcription in EnrichNet but the corresponding score was not significant. The same held for another overlap of 9 genes related to transcription. Overall for this module the significance values were extremely bad. In DAVID the first cluster was related to calcium and cell adhesion, the second to catabolic processes and the third to nucleolus, nuclear lumen and non-membrane-bounded organelles. For all three clusters the enrichment score was only slightly above 1.

Also for the second modules the significance values in EnrichNet

Table 1. Predicted and original cancer stage for the test set based on multiple linear regression and SVM

Prediction Regression	Original	Prediction SVM	Original
0.0	2.0	1.0	2.0
0.0	2.0	3.0	2.0
0.0	2.0	2.0	2.0
0.0	2.0	2.0	2.0
1.0	2.0	5.0	2.0
0.0	1.0	1.0	1.0
0.0	2.0	1.0	2.0
0.0	2.0	2.0	2.0
0.0	2.0	1.0	2.0
0.0	2.0	15.0	2.0
1.0	2.0	1.0	2.0
0.0	2.0	1.0	2.0
0.0	2.0	1.0	1.0
1.0	2.0	1.0	2.0
0.0	2.0	3.0	2.0
0.0	2.0	2.0	2.0
0.0	1.0	5.0	2.0
2.0	2.0	1.0	2.0
0.0	2.0	1.0	2.0
0.0	2.0	14.0	2.0
0.0	2.0	2.0	2.0
0.0	2.0	3.0	2.0
0.0	2.0	3.0	2.0
0.0	2.0	1.0	1.0
0.0	2.0	1.0	1.0

Table 2. Predicted cancer stage for the patients based on multiple linear regression and SVM

Patient	Prediction Regression	Prediction SVM	# Lymph Nodes pos. H&E s. and lm	her2/cent17 ratio
Patient1	0.0	1.0	0.0	1.0
Patient2	0.0	1.0	0.0	1.4
Patient3	0.0	1.0	3.0	1.0
Patient4	1.0	10.0	0.0	11.0
Patient5	0.0	2.0	1.0	1.6
Patient6	0.0	1.0	0.0	1.2
Patient7	0.0	1.0	0.0	1.4
Patient8	0.0	1.0	0.0	1.1
Patient9	0.0	1.0	0.0	1.1
Patient10	0.0	2.0	0.0	2.2
Patient11	0.0	3.0	1.0	3.2

were quite bad. The greatest overlap could be found for the transport pathway and for the small molecule metabolic process. In DAVID the first cluster was related to mRNA and RNA transport, the second to enzyme binding and the third to different diseases like Alzheimer or Huntington. Again, the enrichment score were only slightly above 1 for all three clusters.

The result of EnrichNet for the third modules showed that the maximum overlap was 6 for processes connected to signal transduction but the results were not significant. In DAVID the first cluster was related to inflammatory responses, the second to regulation of synaptic transmissions and the third to the extracellular space. The enrichment score were only slightly above 1.

All in all, the results of the Pearson correlation were not convincing. Another problem is that the mean values of the log2 ratios for the genes in the three modules are almost all close to zero. That means that on average the gene expression for these genes does not differ a lot between normal and diseased.

Partial correlation

Having Applied partial correlation, the resulting network contained 3151 edges. Therefore, around 700 edges were removed applying partial correlation. Nevertheless, the first largest module was the same as for Pearson correlation. Same held for the second largest module. It was the same as for Pearson correlation, too. The only difference was contained in the third largest module. It contained 68 genes and therefore, was larger than for the Pearson correlation

which was unexpected.
In EnrichNet this third largest module had a maximum overlap of 10 with the regulation of transcription and the second largest overlap of nine with transcription. Again, the significance scores were really bad. In David the first cluster was related to regulation of secretion, the second to GTPase regulator activity and the third to leucine-rich repeats. The enrichment scores for all three clusters were slightly above 1.

5.3 Protein expression networks

There have been 42 correlation that were above the given threshold of 0.75, respectively below -0.75 when computing the Pearson correlation. The three largest modules contained 16, 4 and 2 proteins. The network size remained the same when applying the partial correlation. There were still 42 edges in the network. The edges were exactly the same. Therefore, also the resulting modules are the same. That is why only the result of Pearson correlation is considered. Information about the proteins were taken from Gene Ontology (GO , n.d.).
The first protein included in the community is ETS1 which regulates the transcription from RNA polymerase II promoter. ERK2 is a mitogen-activated protein kinase, TIGAR a phosphatase and RBM15 an RNA-binding protein. TAZ is responsible for membrane organization. Rab11 and Rab25 are involved in GTPse activity and therefore, also in signal transduction. PDCD4 is a cell death protein and responsible for cell aging and apoptotic processes. Bap1 deubiquitinates proteins. It is also the protein with the highest degree in this community, it has the highest number of connections to other proteins in this module (see figure 3 E). SF2 regulates gene expression, Cyclin B1 is involved in cell proliferation, SCD in membrane and fatty acid biosynthetic processes and MIG-6, for example, in cell morphogenesis and proteolysis. NF-kB is part of an inflammatory response and PKC-alpha is a protein kinase and part of intracellular signal transductions.
It is not unexpected that these proteins play a role in cell proliferation, transcription or apoptosis. The proteins included in the expression file were already a selection of proteins that are known to be involved in cancer. This preselection also explains why the partial correlation did not remove edges from the network. The selected proteins are known for similar functions and therefore, they are probably also dependent on each other.

6 DISCUSSION

From the results we showed that our invented pipeline works fine with big data sets and is able to quickly build a model based on multiple linear regression or SVM. One can also see that the SVM seems to be better in predicting the cancer stage by evaluating the two parameters we have chosen. This argument can be proofed by comparing the relative prediction error of our two approaches and by looking at the tables 1 of the predicted testing data. Yet both approaches have difficulties in a consequently prediction of the two cancer stages (spreading and non-spreading). That is why one can not really trust our predictions for the patients from the table 2. This could be due to the fact that we generalized the data from four stages (with up to three sub-stages). Our intention was to classify the patient into a spreading and non-spreading type because the future outcome of the cancer and the associated therapy depends heavily on that. But this assumption possibly removed to much information. We also looked at the data with four cancer stages. With that approach the number of lymph nodes positive by H&E staining and lm follows more or less a square root function but the cent2/her17 ratio is a complicated polynomial function. In a future aspect of the pipeline we want to investigate the behavior of the her2/cent17 ratio such that we compare the ratio with different cancer types and look if the ratio is useful for predicting the cancer stage of BRCA. We also tested other parameters but most of them have no functional relationship with the associated cancer stage. Because the multiple linear regression as well as the SVM depends on the dimension and the functional relationship of the data one can also introduce a Kernel based method (approximation).
Afterwards, we tried to figure out which genes and proteins are related to BRCA. Therefore, we built another step which evaluates three gene expression networks and three protein networks based on Pearson and partial correlation. The networks nearly remained the same after applying the partial correlation for each pair found in the gene expressions networks and protein networks based on Pearson correlation. Thus, the correlation in these networks have a high confidence. We expected that this would happen for the protein networks because it seemed that the data sets includes mainly proteins that are related to BRCA or some other cancer types. For the gene expression network we thought that the partial correlation would have a bigger impact. Only the third module changed by gaining more correlations in comparison to the Pearson correlation network. This could be explained by the fact that some genes could be correlated to some other modules based on Pearson correlation. By computing the partial correlation we removed these correlations. Some genes are now stronger correlated to the third module. Hence, we assume that the three largest modules in the correlation networks most likely have a small chance to contain a spurious correlation and that the chances increases for the smaller modules.
We set the threshold for Pearson correlation for the gene expression to 0.9 (-0.9) and for the protein networks to 0.75 (-0.75) to decrease the chance of a spurious correlation. These thresholds were quite high and therefore, it is possible that we excluded correlations that could play a significant role for the BRCA. In a next step we want to repeat the analysis with other thresholds and check if the networks remain the same. It could be done by introducing a bootstrap approach and building a hierarchical model with random forest. This would most likely increase the confidence for the networks and could improve the significance for EnrichNet and DAVID.
The genes that we found in the three biggest modules have a functional relationship that is not surprising for cancer for example the the regulation of the transcription, calcium and cell adhesion and catabolic processes. We found the most interesting clusters in the third module which were related to inflammatory responses and extracellular space. Nevertheless, the main problem was that the significance and enrichment score were really bad for all modules making the results unreliable. For the protein networks we will not investigate other networks based on the data which we used because the proteins are most likely related to cancer or BRCA like ERK2 or NF-kB. But, we want to test if one can use the protein data to build up a new model for classification based on the existing multiple linear regression and SVM to improve the quality for predicting the cancer stage of a patient.

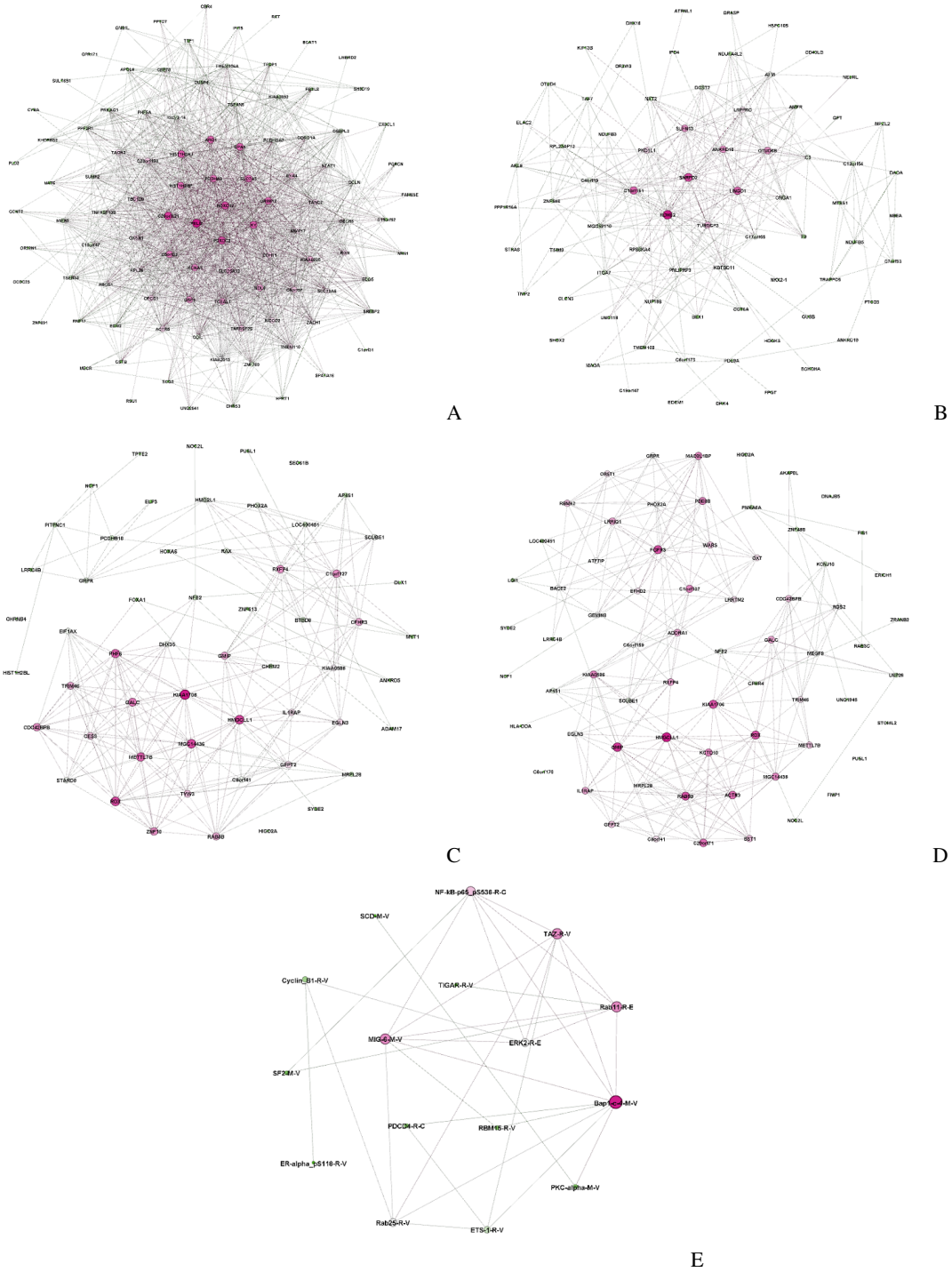

Fig. 3. The resulting modules of the gene expression network an protein expression network analysis were visualized in Gephi (Gephi , n.d.) using Fruchterman Reingold. The size of the nodes is dependent on the degree of the nodes (number of edges). Nodes with a low degree are colored green and nodes with a high degree are colored pink. Picture A shows the largest community of the gene expression network created by Pearson correlation. The nodes are strongly connected, e.g., there are lots of interactions between the genes and it is hard to make out important ones. Three genes, C20orf121, HOXC12 and PELI2 have the most edges. In picture B the secnd largest community is shown. This module is less stronly connected but it has of course also less nodes. The gene RDHE2 has the most connections to other genes in this community. The third module (picture C) is even smaller. KIAA1TDB has the highest degree. The next picture (D) contains the third largest community of the gene expression network that was created by partial correlation. Compared to the third largest module of the Pearson correlation network it has has more node and edges. Picture E shows the result largest community of the protein expression network. The result is the same for Pearson and partial correlation. It only consists of 16 nodes. Bap1 has the highest degree and is connected to 9 other proteins of this community.

7 CONCLUSION

We managed to built a first version of our CNBM (classification and network based method) pipeline. In this version we found out that we generalized the clinical data for the multiple linear regression and SVM too much. Yet, we are still certain that the two parameters - the number of lymph nodes positive by H&E staining and lm and the her2/cent17-ratio - can be used to create a model to classify the cancer stage of a patient. By the results we observed that the number of positive lymph nodes and the her2/cent17 ratio definitely increases with a higher cancer stage. In the future we want to include the protein data in the model. The protein networks had no real value due to the fact that the proteins are clearly pre-selected for BRCA. That is different for the gene expression networks. With our pipeline we found gene expression networks that could play an important role for BRCA. In the future we want to increase the confidence by implementing a bootstrap and random forest approach. Afterwards, we want to go further into detail and look for networks that are present in a specific cancer stage. The cell is no static construct and changes constantly. That makes the investigation for a gene expression network pretty hard but with this approach it is possible to see a possible outcome. The invention of this pipeline was made more difficult by the inconsistence of the clinical data. Therefore, a lot of data has to be removed. That is why we also suggest to integrate a global standard for clinical data.

8 WORK-SHARING

CommunityDetection.java - example copied from https://github.com/-apache/flink/blob/master/flink-staging/flink-gelly/src/main/java/org/-apache/flink/graph/example/CommunityDetection.java
Correlation.java - Irene Ziska except for the partial Correlation that was done by Florian Heyl (begin line 222)
CorrelationProtein.java - Irene Ziska except for the partial Correlation that was done by Florian Heyl (begin line 220)
LargestModules.java - Irene Ziska
AddFunctions.scala - Florian Heyl
Classification.scala - Florian Heyl
Patient.scala - Irene Ziska
ProteinPrepro.scala - Irene Ziska
Regression2.scala - Florian Heyl
MainClass.scala - Florian Heyl

REFERENCES

American Cancer Society, (n.d.), Retrieved August 29, 2015, Types of breast cancers http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-breast-cancer-types#

Cancer Research UK, (n.d.), Retrieved August 29, 2015, Stages of cancer http://www.cancerresearchuk.org/about-cancer/what-is-cancer/stages-of-cancer

Chibon F, de Mascarel I, Sierankowski G. (2009) Prediction of HER2 gene status in Her2 2+ invasive breast cancer: a study of 108 cases comparing ASCO/CAP and FDA recommendations, *An official journal of the United States and Canadian Academy of Pathology, Inc.*, Vol. 22, Num. 3, 403-409, doi:10.1038/modpathol.2008.195, PubMedID: 19060846

Curigliano, G. and Disalvatore, D. and Esposito, A. and Pruneri, G. and Lazzeroni, M. and Guerrieri-Gonzaga, A. and Luini, A. and Orecchia, R. and Goldhirsch, A. and Rotmensz, N. and Bonanni, B. and Viale, G., 2013, Risk of subsequent in situ and invasive breast cancer in human epidermal growth factor receptor 2-positive ductal carcinoma in situ, *Annals of Oncology*, Vol. 26, Num. 4, 682–687, http://annonc.oxfordjournals.org/content/26/4/682.abstract, doi 10.1093/annonc/mdv013

DAVID Bioinformatics Resources 6.7, NIAID/NIH, (n.d.), Retrieved September 5, 2015, https://david.ncifcrf.gov/tools.jsp

Eccles, Suzanne and Aboagye, Eric and Ali, Simak and Anderson, Annie and Armes, Jo and Berditchevski, Fedor and Blaydes, Jeremy and Brennan, Keith and Brown, Nicola and Bryant, Helen and Bundred, Nigel and Burchell, Joy and Campbell, Anna and Carroll, Jason and Clarke, Robert and Coles, Charlotte and Cook, Gary and Cox, Angela and Curtin, Nicola and Dekker, Lodewijk and dos Santos Silva, Isabel and Duffy, Stephen and Easton, Douglas and Eccles, Diana and Edwards, Dylan and Edwards, Joanne and Evans, D and Fenlon, Deborah and Flanagan, James and Foster, Claire, 2013, Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer, *Breast Cancer Research*, Vol. 15, Num. 5, R92, http://breast-cancer-research.com/content/15/5/R92, doi 10.1186/bcr3493, PubMedID 24286369

EnrichNet - Network-based enrichment analysis, (n.d.), Retrieved September 5, 2015, http://www.enrichnet.org/pages/tmp/048c0_1441622976/result.php

Fischer, Andrew H. and Jacobson, Kenneth A. and Rose, Jack and Zeller, Rolf, 2008, Hematoxylin and Eosin Staining of Tissue and Cell Sections, *Cold Spring Harbor Protocols*, Vol. Num. 5, http://cshprotocols.cshlp.org/content/2008/5/pdb.prot4986.abstract, pdb.prot4986, doi:10.1101/pdb.prot4986

Flink Community Detection Example, Retrieved September 5, 2015, https://gitlab.tubit.tu-berlin.de/amr_osman/flinkstacktile/blob/912f8d90efef2912bd9aac7c0b44d05fd90d6803/flink-staging/flink-gelly/src/main/java/org/apache/flink/graph/example/CommunityDetection.java

SVM using CoCoA, (n.d.), Retrieved August 24, 2015, https://ci.apache.org/projects/flink/flink-docs-master/libs/ml/svm.html

FlinkML - Multiple linear regression, (n.d.), Retrieved August 24, 2015, https://ci.apache.org/projects/flink/flink-docs-master/libs/ml/multiple_linear_regression.html#fit

Gephi 0.8 Beta, Retrieved September 5, 2015, http://gephi.github.io/

Gene Ontology Consortium, (n.d.), Retrieved September 5, 2015, http://geneontology.org/

National Cancer Institute, (n.d.), Retrieved August 01, 2015, Breast invasive carcinoma: Case Counts https://tcga-data.nci.nih.gov/tcga/tcgaCancerDetails.jsp?diseaseType=BRCA&diseaseName=Breast\%20invasive\%20carcinoma

Table API - Relational Queries, (n.d.), Retrieved September 5, 2015, https://ci.apache.org/projects/flink/flink-docs-master/libs/table.html

The Cancer Proteome Atlas, n.d.), Retrieved September 5, 2015, http://bioinformatics.mdanderson.org/main/TCPA:Overview