Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Parallelization of a Bioinformatics application with a Workflow Language: A critical analysis of a parallel grid search optimization of the LIMMA algorithm based on the Cuneiform Workflow Language

Exposé of Bachelor's Thesis

eingereicht von: Monika Leung
geboren am: 4.2.1994
geboren in: Berlin

Gutachter/innen: Prof. Dr. Ulf Leser

eingereicht am: .....

1 Background

In cancer genomics, methods are developed to diagnose cancer and to find effective treatments for extending expectancy of life and mitigating symptoms. The ability to compare the gene expression levels of cancer cells and normal cells, allows one to detect differences in the state of cancer cells compared to healthy ones. These differences can be over- or underexpression (i.e. genes seem more or less active) when compared with normal cells. Gene expression can be measured from Messenger RNA (mRNA) for example. Microarrays can be used to compare two to find differentially expressed genes [5].

MRNA-microarray is a technology to measure the expression level of numerous genes simultaneously and to find differentially expressed genes or group of genes of the tested subject. A microarray consists of a collection of probes (single-stranded RNA sequences). The mRNA of the samples is labeled with fluorescent dye and binds through hybridization to the complementary probe on the microarray to form a double-stranded RNA sequence. After hybridization, the microarray is scanned with a laser which results in gene expression data [2]. In this bachelor's thesis, the focus is on gene expression data from mRNA-microarrays.

The gene expression data needs to be analyzed to differentiate between biological signals (expression values) and experimental noise (e.g. from sample preparation or hybridization), in order to obtain the biologically relevant parts of the data [7]. The analysis includes the following steps: background correction, normalization, and test statistics. Background correction removes background noise. Normalization is needed to remove systematic effects caused by technical differences, so that experiments are comparable to each other [2]. Parametric statistical models are used for test statistics. They are used for example to consider the fact that some differentially expressed genes could have been observed by chance.

There are many algorithms that analyze microarray expression data and their methods differ according to the steps yielding different results. For every combination of microarray experiment and algorithm there is a different set of optimal parameters and researchers often do not have the time/resources to traverse the parameter space. Grid searches (sometimes called parameter sweeps) are useful in these cases for finding optimal parameter configurations. The data is run on various algorithms with different parameter configurations. After comparing the results against a target function, optimal choices of algorithms and parameters can be determined. A target function is a measure with which one can assess how good one's results are.

Another issue is that bioinformatics applications require a lot of computation of large amounts of data. Often, the algorithms for gene expression analysis are parallel in concept and since they are used extensively and repeatedly, parallelization can be advantageous [4].

2 Goals

The goal of this bachelor's thesis is to implement a parallelized grid search on the LIMMA differential expression algorithm to analyze how the target functions change in relation to the iteration over input parameters. It will be determined if optimal parameter configurations exist. The term 'optimal' will be specified with respect to different intrinsic target functions (e.g. 'intra- vs. inter-group correlation' or alternatively 'robustness'). The results of the grid search will be compared against an extern gold and silver standard in order to determine if one result is better than another.

3 Approach

3.1 Datasets

This bachelor's thesis works on one gold standard (GSE21344 [8]) and one silver standard dataset (GSE20437 [9]) that come from Gene Expression Omnibus (GEO) [3]. GEO is a public functional genomics data repository which accepts array- and sequence-based data. A gold standard is an experiment where a known amount of RNA is mixed into the sample to create differentially expressed genes when compared to the original sample, so the outcome is known. Whereas a silver standard is an experiment where the outcome is unknown, but the result can be accurately verified to see if it corresponds to reality.

The GSE21344 dataset is a gold standard as it is a controlled spike-in dataset of the drosophila melanogaster genome. Spike-ins are designed to hybridize with a specific probe on a specific microarray and are mixed into the sample in order to measure the degree of hybridization. The GSE20437 dataset is a silver standard. It consists of gene expression data from 4 groups of patients with breast cancer or with a risk of getting the disease. 98 differentially expressed probe sets (86 genes) were found between two subtypes, which were confirmed experimentally in 84% of the cases. The grid search is performed on these datasets, because it is important to assess how close to reality the results of the grid search are.

Since the expression data is raw, a number of normalization methods are performed separately on the data, e.g. RMA (Robust Multi-array Average), GC-RMA (GeneChip Robust Multi-array Average) and quantile normalization. These are part of the parameter set over which the grid search optimizes.

3.2 Grid Search

A grid search is an exhaustive search through a specified set of parameters. With two parameters, it has the structure of a matrix where the rows and columns represent one specific parameter. With each increasing row or column, the value of the corresponding parameter is incremented. Essentially, the cross product of a set of possible values for

each parameter is built. The grid search can also be multidimensional (i.e. more than two parameters). If one perturbs one parameter (e.g. fixates one row and follows the columns) the effect of the changing parameter can be observed in the results. Grid searches can be easily parallelized, since the individual parameter configurations are independent from each other.

3.3 LIMMA

The given algorithm is LIMMA (Linear Models for Microarray Data) [6]. It uses linear models to analyze gene expression data and it can compare many RNA targets simultaneously. LIMMA takes one or two parameters: the design matrix and the contrast matrix, whereas the latter can be omitted in simple experiments. The design matrix indicates which RNA samples have been applied to each array while the contrast matrix specifies which comparisons between RNA samples one wants to make.

The birds-eye perspective workflow of one LIMMA run is as follows: reading the input files or importing expression values, pre-processing the raw data (background correction, normalization), fitting linear models to the data, testing for differential expression, and plotting the results (figure 1).

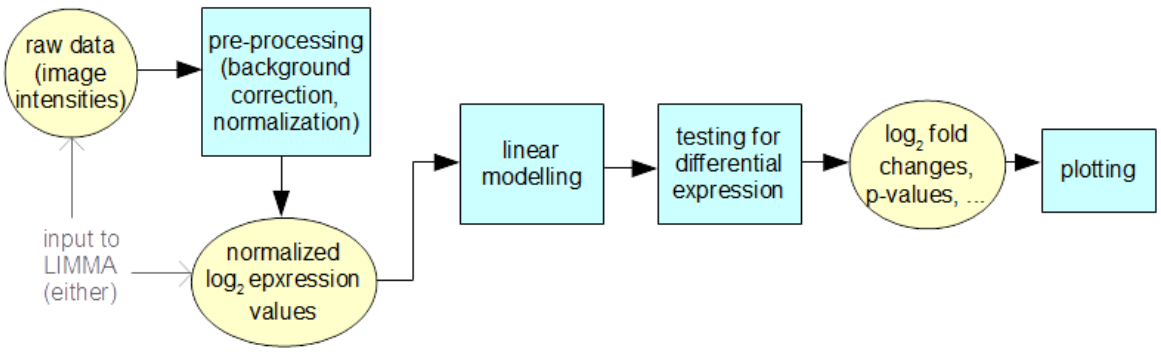

Figure 1: Diagram of the LIMMA workflow.

LIMMA provides functions for reading image output files and it also accepts data objects that contain expression data. The expression values can be normalized log_2-values or image intensities. In the latter case, the data needs to be background corrected and normalized before the analysis. Linear modelling is the core component of LIMMA. It works on a matrix with expression values where the rows contain the genes/probes and the columns contain different samples (biological/technical replicates). Linear models are fitted gene-wise (i.e row-wise) to the gene expression data in order to assess differential expression. LIMMA can fit linear models in two ways, either robustly or by least squares. Comparisons that are of interest can be either estimated directly or given via a contrast matrix. Then, the fitted models and the contrast matrix are taken to compute log_2-fold

changes and t-statistics for the comparisons of interest. The empirical Bayes method (eBayes) is used to test for differential expression, it estimates the prior probability from the data.

The results of the algorithm include log_2-fold changes, standard errors, t-statistics, and p-values and they can be visualized in various plots. In this bachelor's thesis only the fold changes, p-values, overrepresented pathways, and the amount of differentially expressed genes and probe sets are of interest.

3.4 Procedure

The grid search is implemented with the help of the workflow language Cuneiform [1], with which one can implement parallelized scientific workflows. A provided API for LIMMA is used instead of R itself. For handling specific file types, parser scripts need to be written. Before the data is analyzed with LIMMA, it is possibly filtered to remove probes with either low or no expression.

The implementation is written in 3 iterations. The first iteration is implementing the grid search for the two parameters p-value cut-offs and log_2-fold change cut-offs and using RMA normalization for the data. In the second iteration GC-RMA normalization and possibly other parameters (e.g. background correction) are added. Finally, the last iteration is adding quantile normalization and possibly other parameters (e.g. eBayes modes). At the end of every iteration the run time is evaluated and the results are compared to the gold and silver standard (figure 2).

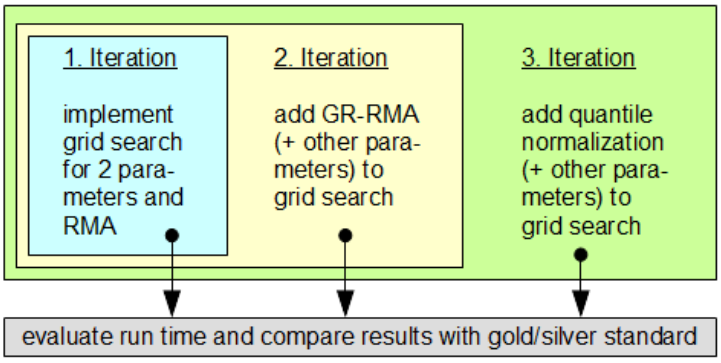

Figure 2: Diagram of the iterations for the grid search implementation.

After the implementation of the grid search is completed, the run time, the RAM usage, and the parameters' effect on the results are observed. Next, the various parameter configurations are evaluated by relating target functions to the target function of the distance to the gold or silver standard (i.e. the number of differentially expressed genes not found in own analysis) to determine which configurations are optimal (figure 3). The final step is to compare the run time of the parallelized implementation against the run

time of the original implementation (which is not run sequentially but the run time of one run is extrapolated) to assess the advantages and disadvantages of parallelization.

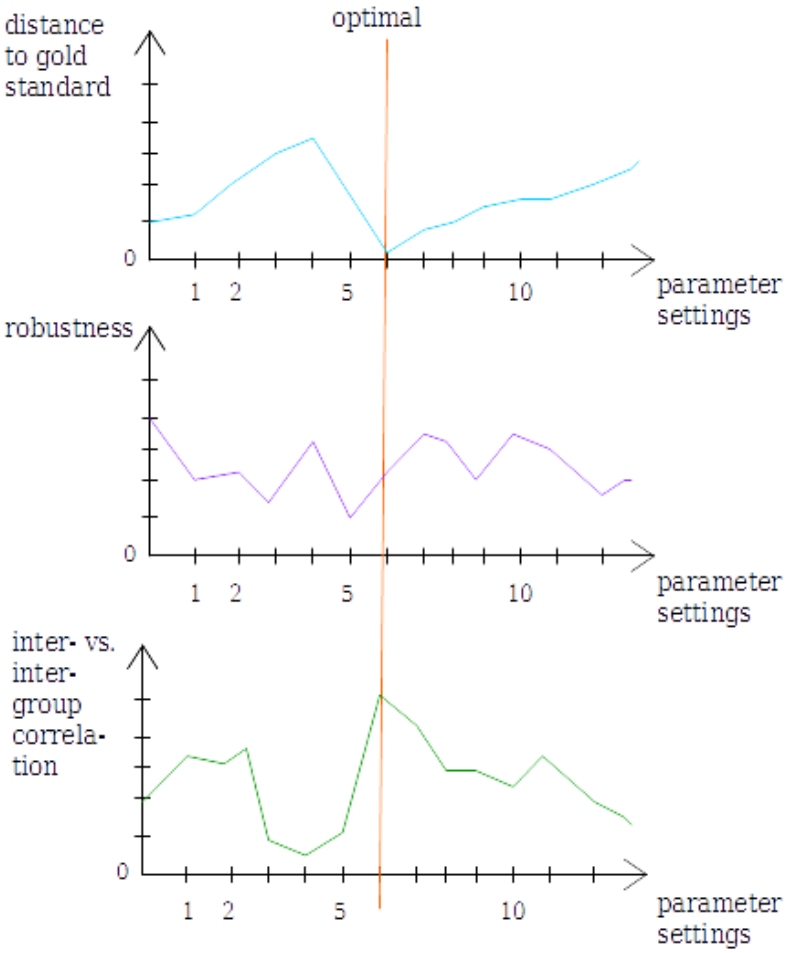

Figure 3: Example illustration of three target functions and an optimal parameter setting. This is an example illustration with fictitious values. The diagrams show how a suitable target function for the grid search can be identified and how the optimal parameter setting can be found. The optimum is at the parameter setting where the distance to the gold standard is minimal (here setting nr. 6). The correlation function seems to be a suitable target function as the maximal correlation coincides with the minimal distance, whereas the robustness function seems rather random here.

References

[1] J. Brandt, M. Bux, and U. Leser. "A functional language for large scale scientific data analysis." *BeyongMR, ICDT/EDBT Workshop*, 2015

[2] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments", *Statistica sinica*, 111-139, 2002

[3] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", *Nucleic Acids Res.*, 30(1):207-10, 2002

[4] M. Gaggero, S. Leo, S. Manca, F. Santoni, O. Schiaratura, and G. Zanetti, "Parallelizing bioinformatics applications with MapReduce", Cloud Computing and Its Applications, p.22-23, 2008

[5] J. Khan, L. H. Saal, M. L. Bittner, Y. Chen, J. M. Trent, and P. S. Meltzer, "Expression profiling in cancer using cDNA microarrays", *Electrophoresis*, 20(2), 223-229, 1999

[6] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies", *Nucleic Acids Res.* 43(7), e47, 2015

[7] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments", *Proc. Natl. Acad. Sci. USA*, 99(22), 14031-14036, 2002

[8] `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21344`

[9] `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20437`