Efficient Algorithm for Microarray Probes Re-annotation

Pawel Foszner, Aleksandra Gruca, Andrzej Polanski,
Michal Marczyk, Roman Jaksik, and Joanna Polanska

Silesian University of Technology, Institute of Informatics,
Akademicka 16, 44-100 Gliwice, Poland
{pawel.foszner,aleksandra.gruca,andrzej.polanski,michal.marczyk,roman.jaksik,joanna.polanska}@polsl.pl
http://www.polsl.pl

Abstract. Systems for re-annotations of DNA microarray data for supporting analysis of results of DNA microarray experiments are becoming important elements of bioinformatics aspects of gene expression based studies [10]. However, due to the computational problems related to the whole genome browsing projects, available services and data for re-annotation of microarray probes are still quite sparse. The difficulty in developing systems of re-annotations of microarray probe sets is mainly in huge sizes of probe set data.

In our research we have created an efficient re-annotation method by combining the well known gene search tool BLAT with appropriately designed database. The elaborated system extends possibilities of existing re-annotation tools in the sense that: (1) by tuning parameters of all steps of re-annotation procedure any Affymetrix microarray standard chip can be automatically re-annotated in few hours, (2) several Affymetrix microarray chip standards are already precomputed.

Keywords: re-annotation, microarrays, expression data, affymetrix, classification.

1 Introduction

In the Affymetrix microarrays, the gene intensity is estimated/calculated on the basis of signals obtained from gene probes consisting of 25-mer oligo-nucleotides. Procedures for merging and normalizing signals from probes aim at obtaining reliable estimates of values of expressions of genes. In many cases, however, the estimated value does not match the true value of the gene expression. Technical sources of measurement noise introduce random error with controlled amount of variation. Important source of the mismatch between estimated and true values of expression is in the design procedure of microarray probes. One source of the design error is the presence of single nucleotide polymorphisms inside the oligo - sequence of the chip. More important errors coming from the design procedure of the microarray are related to assigning the probe to the locus in the genome

P. Jędrzejowicz et al. (Eds.): ICCCI 2011, Part II, LNCS 6923, pp. 281–289, 2011.

different than desired. These errors are related to imperfections of assembly processes of genomes of organisms, like false sequences, gaps etc. The process of assembly of genomes is, however, continued and successively improved assemblies of genomes are published. Successive improvement of the quality of available data on genomic sequences opens the possibility of improving the quality of microarray measurements by re-annotating microarray probes, i.e., re-targeting all microarray probe sequences to newest versions of genomic sequences and modifying microarray definition files (CDF files) according to the results of this procedure. Researches focused on such possibility were already published in the literature, e.g., [10].

In this paper we report results of our work on creating and implementing an efficient algorithm for accurate targeting strings of nucleotide sequences, coming from microarray probes, to genomic databases. We have developed a methodology which provides information about how the probes are aligned to the latest built of the human genome. We used publicly available tools and databases for genomic sequences, BLAT [8], EST, mRNA, RefSeq databases [7] and the UniGene database [3]. Using mRNA and EST sequences databases we searched for matches to the genome, for each probe of the DNA microarray.

We have embedded the databases and technologies in the PHP and MySQL environment. The elaborated methodologies and tools allow the user both to access the re-annotation tools as a standalone program and as a web service. Our aim was also to use the elaborated tools for studying the improvement in gene probes definition obtained in successive assemblies of genomes, namely we researched statistics of status of gene probes. Status of the gene probe is determined by results of alignment of the gene probe against genomic databases, which enable verifying whether the gene probe (1) correctly and uniquely targets the desired gene, (2) miss-targets the gene, (3) targets the desired gene but does not fulfill the uniqueness property [10].

Results of using our algorithms for Affymetrix DNA microarrays are presented at two levels. The higher lever involves statistics stemming from comparisons of the status of original and re-annotated probes and the lower one involves comparing estimates of expression profiles, for exemplary data, between original and re-annotated microarrays. Contemplating the results reported in our research one can check how incorporating the latest improvements in human DNA assembly in DNA microarra data processing algorithms, can improve estimation of values of gene expressions.

2 Genomic Databases

All collected data concern human genome version 19. The entire human genome in this version has been downloaded from the UCSC website (University of California Santa Cruz) in 2bit format [7]. A .2bit file stores multiple DNA sequences, where each nucleotide is stored on 2 bits. This file was a database used by the BLAT program [8] for searching for matches of microarray probes. All the matches found by BLAT were linked to EST and mRNA sequences and further

assigned to corresponding genes. Therefore, there was a need to have a database of EST and mRNA sequences and database with information about the assignment of these sequences to genes. Databases that include sequences were downloaded from UCSC website [7] (files all_mrna.txt.gz and all_est.txt.gz). Database that contains information about assignment a sequence to the gene, was UniGene database. Downloaded version was 228, from the NCBI servers [3]. The last part of the algorithm verifies if the aligned probe does not belong to the non-coding region.

2.1 Pre-processing

Data obtained from the UCSC, were flat files representing MySQL tables. These are very large tables without any indexes, and without any possibility of creation of any unique index. This causes that the simplest search on a table with millions of rows becomes very costly in time. Given that microarrays have hundreds of thousands of probes, this results in unacceptable execution time.

BLAT [8] as a result of the calculation returns the name of the chromosome on which the match was found, and the starting and ending index where the match was found. Thus, in the SELECT statement we had two variables of type numeric and one string. Elimination of this last variable was a key element of the time optimization process. The main table has been divided into many tables, where every single table contains data on one chromosome. This resulted in the elimination of a string variable from select statement, which accelerated its execution.

At each stage of the algorithm, the structure of the database was changed for the most expensive queries. This fragmentation also allowed us to create multiple unique indexes, which again contributed to the time optimization.

3 Implementation

As a first step, we search for matching sequences in the human genome using the BLAT program. One of the pre-processing steps which BLAT program performs, is creation of a map of indexes for the searched database. This map is placed entirely in computer memory, and all search operations are performed on it. The map of the entire human genome build 19 uses around 6GB of memory, thus it was necessary to recompile BLAT program for use on x64 system architecture. The results of the program after compilation has been checked for consistency with the 32-bit version.

BLAT call parameter were default parameters for the standalone version, optimized for short sequences with maximum sensitivity [7] that is:

```
blat -stepSize=5 -repMatch=1000000 -minScore=0 -minIdentity=0 19.2bit
query.fa output.psl
```

The above parameters have the following interpretations: 19.2bit is a file including the data of the human genome build 19. The next file is a specially prepared FASTA file with probes sequences. It includes all the probes and each

of them is given a unique name. This name consists of a symbol of a probeset to which probe belongs and the coordinates defining its position on array. Below we present beginning of the file used for the re-annotation array HG-U133A:

```
> 1007_s_at:467:181
CACCCAGCTGGTCCTGTGGATGGGA
> 1007_s_at:531:299
GCCCCACTGGACAACACTGATTCCT
> 1007_s_at:86:557
TGGACCCCACTGGCTGAGAATCTGG
```

The last call parameter is the file which stores the results. After the execution of the BLAT program, we save only such results, which have no more than two mismatches and we remove all other probes from further analysis.

As the input the system takes results of the BLAT program and a list of probes with their unique names used in the FASTA file. In the next step, for all of the matches, we search for EST and mRNA sequences. Therefore, for each match we obtain a list of sequences that may belong to a gene. Annotation of a sequence to a particular gene is verified using the UniGene database and we keep only those that represent a gene according to the UniGene database. The last step of the algorithm verifies whether found match is in the coding or non-coding region.

The result of the program is a report including information about analyzed sequences. For each sequence we provide information about the matching gene, and whether it represents either a single or many genes. In addition, there is an information whether the probe is located in the coding or non-coding region.

Finally, we create a CDF file which includes only those probes that represent single genes and are located in the coding regions of these genes.

4 Re-annotation

To verify the quality of the performed re-annotation, we analyzed how it affects the expression data after normalization process. For this purpose, we compared expression values computed for two different CDF files. The CDF file describes the layout for an Affymetrix GeneChip array [1]. First of them was the original Affymetrix CDF file downloaded from the official web site and it was used as a reference file. The second one was generated on the basis of our method. This example re-annotation and CDF files were prepared for the HG-U133A array.

In the table 1 we present some statistic for the the CDF files for the HG-U133A matrix. The first column of the table includes statistics for the original Affymetrix CDF file and the second one represents the results after our re-annotation algorithm.

4.1 Microarray Data

The data were downloaded from the NCBI GEO DataSets database [3]. The example set has accession number GSE10072. This set contains data for the set

Table 1. A table showing the array HG-U133A statistics

	Affymetrix CDF file	Our CDF file
Number of probesets	22 216	11 708
Number of probes	22216	10954
Unigene build	133	228
Unigene release date	April 20, 2001	October 1, 2010
Human genome build	31	37.1

of samples of human tissues classified as lung cancer tissues or healthy tissues, described in the paper [9]. The study includes 107 tissues (58 tumor and 49 non-tumor tissues) from 74 patients. Each microarray experiment (tissue sample) is stored as a separate CEL file.

4.2 Microarray Expression Normalization

The raw data files (CEL files), were normalized using RMAExpress (Version: 1.0.5 Release) [4]. Normalization parameters were assumed as their default values, which are:

```
- Background adjust = yes
- Normalization = Quantile
- Summarization Method = Median Polish
- Store Residuals = false
```

After normalization we obtained two text files containing normalized expression measurements. Each of these files included expression values computed using different CDF file.

Further analysis were performed for three different expression groups: (1) the entire data set, (2) healthy tissues and (3) tumor tissues. The expression files were loaded into Matlab [2], and expression measurements were averaged in each of the group separately. Finally, we obtained three scatter plots, each for the different expression group.

4.3 Results of Expression Analysis

To analyze if, and how the values of expression levels differ before and after re-annotation, we created scatter plots, where each point represents normalized expression values. During the process of re-annotation we changed an arrangement of the probe sets, and created a completely new, based on alignment a probes to genes. Probes matched to the same gene were grouped into sets.

The plots in figures below show, for each gene, comparison the level of expression of the same data normalized by two different annotation methods. We took into account only the genes common to both CDF files. In our CDF file each gene is represented by a single probeset as opposed to original CDF file, where several probesets can represent one gene. Annotation of original Affymetrix probesets to

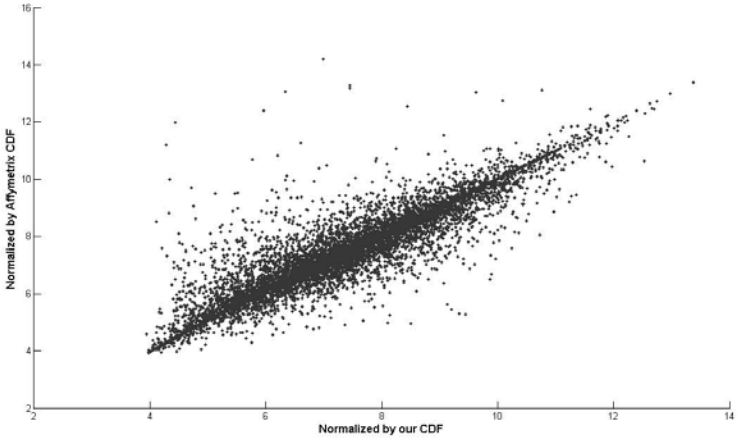

Fig. 1. Scatter plot for all 107 tissues

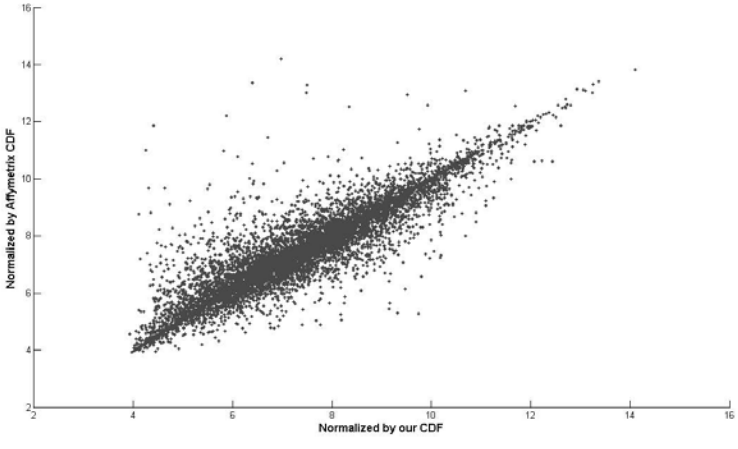

Fig. 2. Scatter plot for subset of normal tissues

genes were taken from HG-U133A Technical Support Documentation on official web site. We used NetAffx Annotation file, release 31 (August 23, 2010).

Scatter-plots were calculated for three subsets of data consisting of 107 microarray experiments. Figure 1 represents analysis of expression values for the entire data set. For each probeset, we averaged expression levels, and after that we compare this averaged values between data normalized using different CDF files. Figures 2 and 3 presents the expression values calculated for healthy and cancer 3 tissues respectively.

We also created two scatter-plots representing expression values for normal and tumor tissues. Fig 4 represents expression values obtained for original Affymetrix CDF files while the figure 5 shows the expression values computed for our CDF file.

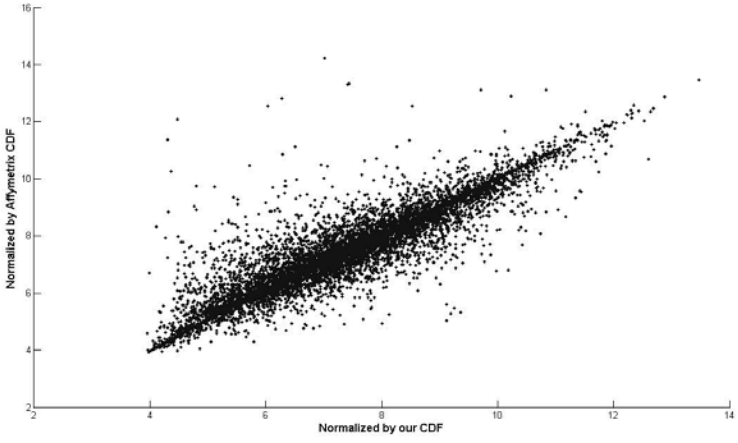

Fig. 3. Scatter plot for subset of tumor tissues

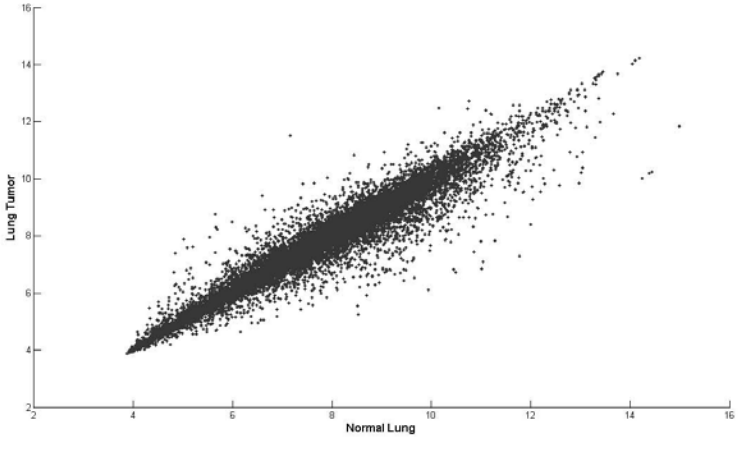

Fig. 4. Tumor tissues vs normal tissues – expression values obtained for original Affymetrix CDF file

Analysis of the results presented in figures 1 – 3 shows that expression values obtained using our method differs significantly from the expression values computed for original Affymetrix CDF file. By analyzing presented figures we can also observe that the differences between annotation methods are bigger than differences between normal and tumor tissues presented in figures 4 and 5. Such differences may affect the results of classification, clustering or many other operations performed on microarray expression data.

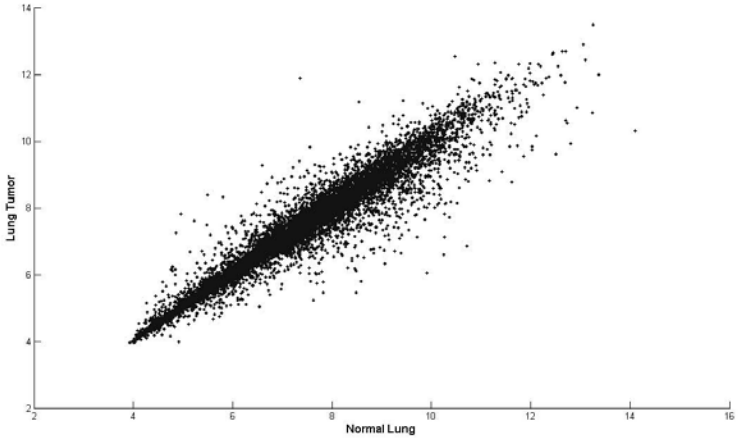

Fig. 5. Tumor tissues vs normal tissues – expression values obtained for our CDF file

5 Conclusions

In this paper we have presented new efficient re-annotation algorithm for Affymetrix DNA microarray probes. By using our algorithm we are able to target all probes of a microarray (of numbers of orders 100,000 - 800,000) to their true genomic location. Then, by reading annotations of their matches we can redesign procedures for computing values of expressions of genes of the microarray. The algorithm uses several available and widely used tools for genomic alignments, the most important one is the BLAT system. The algorithm also links to appropriate genomic databases such as EST, mRNA, RefSeq. These tools and services are combined with fast data parsing system and a local database. The project of the local database is optimized with respect to computational time.

Consistently to already published papers [10] our results show substantial differences between original projects of DNA microarray chips (as an example Affymetrix chip HGU133A is used) and the true locations of probesets. These results are shown in table 1. Moreover, when logarithms 2 of expression values are computed, by using standard RMA normalization procedure, for original project of the chip and for the chip with re-annotated probesets then again substantial differences are observed. These differences are presented, as scatter-plots, for real DNA expression data in figures 1 – 3. Importantly, by visual comparisons of figures 1 – 3 with the scatter-plots in figures 4 and 5, where data on healthy tissues for samples from [9] are compared with data on cancer tissues, stronger differences are observed for re-annotation effect than for the effect of difference between cancer and normal.

6 Further Research

In the further research we plan to re-annotate more Affymetrix arrays for different microarrays and organisms. We will also compare our result with normalized expression data obtained using different re-annotation methods, for example Ferrari [6] or Dai [5].

We also plan to analyze how the re-annotation can affect the quality of classification by comparing misclassification rates for classification of microarray data obtained using the official affymetrix CDF files and CDF files created by us. The information obtained from re-annotations reflects our current biological knowledge and thus application of updated CDF files can significantly improve the classification results.

Acknowledgments. This work was supported by the European Community from the European Social Fund.

References

1. Affymetrix official website, `http://www.affymetrix.com`
2. Matlab - the language of technical computing, `http://www.mathworks.com/products/matlab/`
3. National center for biotechnology information, `http://www.ncbi.nlm.nih.gov`
4. Bolstad, B.M.: RMAExpress Users Guide. 1.0.5 Release (2010), `http://rmaexpress.bmbolstad.com/`
5. Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G.: Evolving gene/transcript definitions significantly alter the interpretation of genechip data. Nucleic Acids Res. 33, e175 (2005)
6. Ferrari, F., Bortoluzzi, S., Coppe, A., Sirota, A., Safran, M., Shmoish, M.: Novel definition files for human genechips based on geneannot. BMC Bioinformatics 8, 446 (2007)
7. Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T.R., Giardine, B.M., Harte, R.A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R.M., Learned, K., Li, C.H., Meyer, L.R., Pohl, A., Raney, B.J., Rosenbloom, K.R., Smith, K.E., Haussler, D., Kent, W.J.: The ucsc genome browser database: update 2011. Nucleic Acids Research 38, D876–D882 (2010)
8. Kent, W.J.: Blat-the blast-like alignment tool. Genome Res. 12, 656–664 (2010)
9. Landi, M.T., Dracheva, T., Rotunno, M., Figueroa, J.D.: Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS One 3(2), e1651 (2008)
10. Nurtdinov, R.M., Mikhail, O.V., Ershova, A.S., Lossev, S.I., Karyagina, A.S.: Plandbaffy: probe-level annotation database for affymetrix expression microarrays. Nucleic Acids Research 38, D726–D730 (2010)