Integrating data from heterogeneous DNA microarray platforms

Eduardo Valente[1*] and Miguel Rocha[2]

[1]Computer Science and Information Systems, Polytechnic Institute of Castelo Branco, Castelo Branco, Portugal, http://www.ipcb.pt

[2]Centre of Biological Engineering, University of Minho, Braga, Portugal, http://www.uminho.pt

Summary

DNA microarrays are one of the most used technologies for gene expression measurement. However, there are several distinct microarray platforms, from different manufacturers, each with its own measurement protocol, resulting in data that can hardly be compared or directly integrated. Data integration from multiple sources aims to improve the assertiveness of statistical tests, reducing the data dimensionality problem. The integration of heterogeneous DNA microarray platforms comprehends a set of tasks that range from the re-annotation of the features used on gene expression, to data normalization and batch effect elimination. In this work, a complete methodology for gene expression data integration and application is proposed, which comprehends a transcript-based re-annotation process and several methods for batch effect attenuation. The integrated data will be used to select the best feature set and learning algorithm for a brain tumor classification case study. The integration will consider data from heterogeneous Agilent and Affymetrix platforms, collected from public gene expression databases, such as The Cancer Genome Atlas and Gene Expression Omnibus.

1 Introduction

DNA microarrays have been widely used in the study of several organisms and phenotypes, with applications ranging from biological sciences to health care and biotechnology. One recent application has been in the study of different types of cancer [1]. The correct classification of the pathology of a patient, in particular in cancer, is essential to decide which drugs or therapies may be applied [2, Chapter 2][3][4]. Often, tumor samples with an atypical morphology complicate the analysis. In addition, certain types or subtypes of tumors may have very little differentiation between them [5].

The analysis based on gene expression is extremely important, given the gaps that traditional diagnosis methods still present. These data are obtained by measuring the amounts of mRNA in a sample for the different genes in study. DNA microarrays can monitor simultaneously expression profiles from a large number of genes [6], as each microarray slide can carry a high amount of probes. The major issue for statistical microarray data analysis is dimensionality. In

*To whom correspondence should be addressed. Email: eduardo@ipcb.pt

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

a typical experiment, a table with thousands of genes for a small number of samples is obtained. This leads to situations where it is difficult, or even impossible, to employ classical prediction algorithms leading to poor prediction accuracy in discriminant models [7].

While a great number of studies have been focused on data dimensionality reduction [8, Chapter 13], applying statistical methods like Principal Component Analysis or Linear Discriminant Analysis, other works turned their attention to the increase in the number of available samples [9]. One way to increase sample size is the integration of microarray data from different studies over the same phenotypes. When the platforms have the same identifiers, this implies a data adjustment to minimize batch effect, understood as non-biological experimental variation across multiple microarray experiments [10]. When the platforms have different feature identifiers (i.e. different probesets), a re-annotation method is needed to make a bridge between them.

Here, it will be presented a re-annotation process based on transcripts, aiming to achieve a universal process that allows the integration of data from distinct microarray platforms. The re-annotated data will then be integrated to form a richer dataset, using different approaches, focusing on reducing batch effect. The resulting integration will be applied on brain tumor grade classification, performing feature selection and classification through filter and wrapper methods. The final result is a methodology and a set of computational tools that allows researchers on this area to integrate several gene expression datasets and apply machine learning algorithms over them, obtaining a subset of the best features-algorithm set for a specific case study. With the re-annotation based on transcripts it is possible to cross statistical significance with biological relevance and find the best subset of biomarkers.

2 Microarray fundamentals

Gene expression values are obtained by measuring amounts of mRNA in samples. Early methods targeted one gene (or a small number of genes) at a time requiring an a priori hypothesis suggesting which gene could be of interest. This limitation has been surpassed with the rising of DNA microarrays. Microarrays are physical structures, which have coupled thousands of specific DNA probes in a plate with a diameter of less than 250 microns [4]. The sequences of these DNA probes, when in contact with a sample, will hybridize with the complementary mRNA segments.

2.1 Affymetrix probesets

In single channel microarrays, the measurement corresponds to a certain amount of expression, which is the case for Affymetrix® platforms. These probesets contain 11 or 16 25-mer probes, each probe measuring a specific zone of the gene. The small size of the sequence means that there is a great chance of cross-hybridization in these probesets. Each probe of Affymetrix is actually a pair of probes. One of these probes is the exact complementary sequence of the target to measure, called Perfect Match (PM), and the other has only a change of one base in the middle of the string, called Mismatch (MM) [11][12]. This change in the basis of MM

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

probes is considered sufficient to not hybridize with the target, so it is considered that this probe measures background noise. The specific intensity for a probe is thus given by Eq. 1.

$$Probe\,Signal = PM - MM \tag{1}$$

To calculate the probeset expression, Affymetrix statistical methods ($Affy_Alg$) are applied over the *n* probes of the probeset, which aim, among other purposes, to exclude from the calculation the pairs of probes that present levels outside the normal parameters [13]. The final value of a probeset measurement is often a log_2 intensity (Eq. 2).

$$Probeset\,Signal = log_2(Affy_Alg(nProbe\,Signal)) \tag{2}$$

2.2 Agilent probesets

When a microarray considers the existence of two samples on the same experiment, it is called a two channel microarray. That is the case with many Agilent® platforms. An Agilent probeset is composed of several identical probes, with a target of 60-mers oligonucleotides. The measurement of the signal by spot (probe) is made using an image processing that measures the average intensity of pixels of this spot [14]. The datasets used on this work use a cell line, established through reference samples of various tissues, which serves to ameliorate the errors inherent in the platform, like for example the background error. The cell line is marked with green dye (Cy3) and the target sample is marked with red dye (Cy5). The signal that is returned is a logarithm of the ratio (Eq. 3).

$$Probeset\,Signal = log_2(Cy5/Cy3) \tag{3}$$

A microarray experiment generally results in an extensive list of probesets and their expression intensities (single-channel) or intensity ratios (dual-channel) [15]. With the current state of the art, it is not possible to cover all genome with probes. There are several zones of the genes that are not monitored on a microarray experiment. This means that the expression of some exons is not measured and, thus, the ”gene expression” concept is actually flawed. The target of a probeset is, in fact, a transcript and not a gene. To meet this proposal, probe designers try to direct targets to exons that belong to a single transcript [16]. Another goal when designing microarray probes is specificity. The genome areas that present high levels of repeated patterns are discarded, because these areas tend to repeat all over the genome and this would cause cross-hybridization.

3 Related work

There have been several studies targeting data integration between microarray platforms. A few are summarized in Table 1. The main distinction is at which level the interface between

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

platforms is made. Since the majority of the datasets available are at the probeset level, the first approach could be to directly map probeset values. However, the data files generally do not have sufficient information to achieve this mapping. The most traditional method is to integrate at the gene level, where it is possible to resort to probeset-gene mapping files available from manufacturers. However, several inconsistencies were pointed out for this approach and recent studies are turning to the transcript level [17].

Culhane et al. [18] made a co-inertia analysis to relate datasets without the need of identifiers annotation, using a multivariate method that identifies co-relationships in multiple datasets. The method was able to visually cluster genes with similar expression patterns between platforms. The authors claim that it is possible to assist in the selection of the strongest features from each dataset for subsequent analysis. This approach prevents the bottleneck caused by the annotation based methods for integration, that limit the genes on study to those identified in all the datasets involved.

Woo et al. [19] compared the variability in expression between three different types of microarrays (Affymetrix GeneChip, cDNA and oligonucleotide). The gene identifiers and probeset-gene mapping were obtained from TIGR and GenBank. The expression value for each gene was obtained by averaging probesets linked to the gene, adding experimental effects like dye and array effects. The platforms were compared by concordance of the F-statistics using a 2x2 cross-classification of significant vs. non-significant genes. Affymetrix microarrays shown greater variation in the magnitude of expression between replicates. However, in terms of statistical significance, a greater concordance between Affymetrix GeneChips and oligonucleotide arrays was shown. The analysis was done at the gene level.

Jarvinen et al. [20] compared data between different expression microarrays, using Affymetrix, Agilent cDNA and a custom-made microarray from a cDNA library. The comparison was made at the gene-level, using the UniGene cluster ID as a common handle among platforms. The expression values of probesets with the same ID, called clones, were averaged. A total of 7936 identifiers for Affymetrix, 7117 for Agilent and 7273 in the cDNA microarray were identified, but only 2340 identifiers were in common among the three platforms. After processing information, the authors obtained 1147 identifiers in common, to which they applied the Pearson and Spearman correlations to compare the gene expression between platforms. The results indicate variability across the three platforms, being greater among commercial arrays.

Ballester et al. [11] established an annotation pipeline at the transcript-level. The probe sequences were collected from the Affymetrix site and were aligned with the genome using the Ensembl exonerate tool. An association probeset-transcript was made if at least 50% of their probes match the transcript, considering 1 bp as maximum mismatch. Probes with match in more than 100 different places on the genome were discarded. Aiming to improve annotation, the 3' untranslated regions (3' UTRs) were considered.

Bellis [21] pointed the problem of annotation dependency on the reliability of files provided by manufacturers with the association of probesets-transcripts-genes. The authors refer that the assumption that an exon read leads to a specific transcript is ambiguous, which implies the need for extra validation. To analyze if two probesets annotated to the same gene are measuring the same entity, they used Pearson correlations, complemented by networks of positive and negative

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

correlations. This structure compares all the probesets and genes and assigns them to different classes. The result was a full textual description of the association of each probeset to genes, exons and transcripts.

Table 1: Related works

Level	Platforms	Reannotation	Data integration	Feature selection	Work
Probeset	Oligonucleotide; Spotted cDNA	clustering	no need	direct from clustering evidences	[18]
Gene	Spotted oligonucleotide; Spotted cDNA; GeneChips	open databases	expression average	gene rank	[19]
	Oligonucleotide; Spotted cDNA; Custom cDNA	open databases	expression average	expression correlation	[20]
Transcript	Oligonucleotide	genome alignment	not considered	not intended to	[11]
	Oligonucleotide	annotation files	average of clones with correlation	not intended to	[21]

Integrating multi-platform microarray data raises many difficulties, and therefore it is often avoided due to the lack of good results. Other authors who have taken this challenge focus on specific integration issues, and the methods used widely diverge. Kostadinova [22] examined how the combination of several related microarray datasets affects different areas of preprocessing and analysis of gene expression data, such as missing value imputation, gene clustering and biomarkers detection. Meng et al. [23] used a top-level approach, with multiple co-inertia analysis to relate two datasets, by maximizing the covariance between eigenvectors in paired dataset analysis. With this method, they can integrate and compare multi-omics data, independent of data annotation. The majority of the authors also focus on a single vendor or on datasets that are only one-channel or two-channel based. Papiez et al. [24] combined data sets from two types of microarrays: oligonucleotide and cDNA. They extract a set of genes common for both platforms and remove batch effects obtaining combined p-values from the two experiments. Even being different platforms, both datasets considered only intensity values.

Although being difficult, data integration has been seen as an important achievement, even leading to proprietary solutions, like the ones described by Willis et al. [25]. In this work, a combination of Thomson Reuters solutions was used, including the Data Annotation and Processing tool, MetaCore Pathway Analysis and the Biomarkers Module of Thomson Reuters Integrity for a simple workflow to process omics data and identify biomarkers of target engagement through pathway analysis and data integration.

4 Methodology

This work intends to establish a general methodology/architecture for microarray data integration and application. The process begins with data collection of gene expression datasets from different free available sources. These datasets must be normalized for expression values and phenotypes, to achieve a common structure across them. The re-annotation processes are then applied to bring heterogeneous annotation datasets into a common backbone of features. The

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

datasets could then be integrated, taking into account the batch effect, and applied on a case study where it is possible to select the best features and classification algorithm. The system was named Integrated Gene Expression Information System (IGEIS) (Fig. 1).

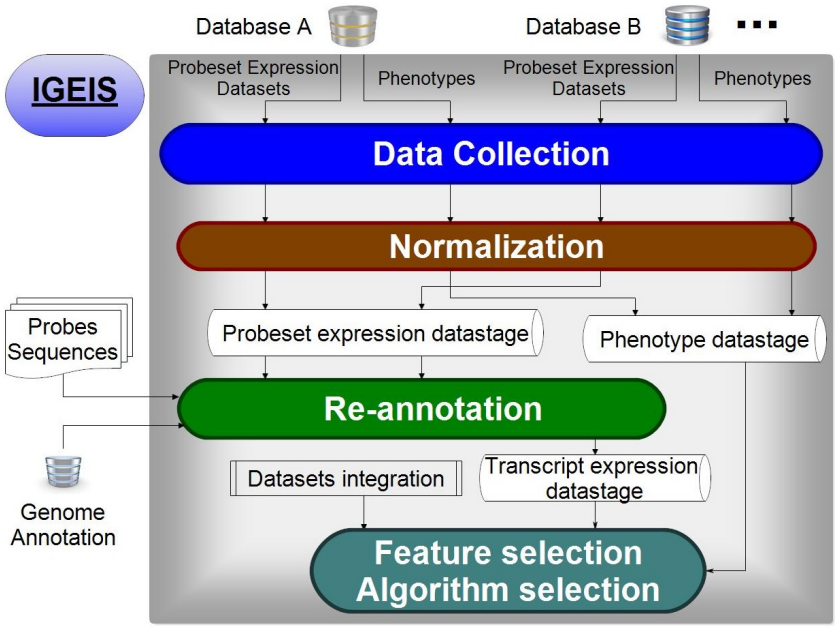

Figure 1: IGEIS structure

4.1 Data collection

The first decision to be made is at what level to perform the data integration: at microarray spot level, probeset level, transcript level or at the gene level. An important factor is at which level the data are available. If data are on a higher level, then it is not possible to descend to a lower level. This stage also encompasses scaling of expression values, as some are on a logarithmic scale and others are on an absolute scale. As most differential expression tools use the logarithmic scale, this was chosen. A more intricate problem is the standardization of phenotypic meta-data. Clinical data are often provided in plain text. This leads to a need of manual intervention to achieve homogeneous and unambiguous information that associates gene expression levels to specific phenotype attributes.

4.2 Re-annotation

The consistent homogeneous feature annotation is a very important step to compare expression data from different platforms. It is only possible to combine values of expression of a probeset with another one if both have the same target. The manufacturers provide annotation to their probesets, but with the constant evolution of platforms, this is quickly outdated [11]. The probeset targets are typically (sets of) exons that comprise transcripts. If two probesets have different transcripts as target from the same gene, it is not surprising that they present different expression values.

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

The re-annotation starts with probe sequence alignment with the genome. This stage is done by collecting the sequences of each probe and registering all the matches those sequences present along the genome. As only exons are expressed, the matches of interest are the ones that lie on these sections. Given the exon-transcript correspondence, it is then possible to establish a probe-transcript correspondence, having in mind that the same exon can belong to multiple transcripts due to alternative splicing. For the contribution ratio of a probeset to a transcript expression value, the quantity of matches that the probeset makes with the transcript will be used, divided by the total matches on the whole genome (Eq. 4).

$$psetR_{(ij)} = \frac{npm_{(ij)}}{npm_{(ij)} + npn_{(i)}} \tag{4}$$

where: *psetR* - probeset ratio for this transcript; *npm* - number of probe matches on the transcript; *npn* - number of probe matches outside the transcript; *i* - probeset index; *j* - transcript index.

For the determination of *npn*, it is necessary to check if each hit is outside the bounds of the target transcript, because a probe may have a hit in another transcript that has overlapping sections and, therefore, be identified erroneously as an outside match. Where it is not possible to find the probe sequences, or when a probe does not have matches, it will be discarded. A transcript expression value will be the weighted average of corresponding probesets (Eq. 5):

$$te_{(j)} = \frac{\sum_{i=1}^{n}(psetV_{(ij)} * psetR_{(ij)})}{\sum_{i=1}^{n} psetR_{(ij)}} \tag{5}$$

where: *te* - transcript expression; *psetV* - probeset expression value; *n* - number of probesets that match transcript j. With Eq. 5 it is possible to restructure datasets on a transcript based manner.

4.3 Datasets integration

The method used for integration may be generic or adapted for a specific case study. On brain tumor grade determination, the problem to solve is a classification task, with 4 distinct classes (1 to 4), being 1 the less malignant grade and 4 the higher malignant grade, called glioblastoma. To exemplify, the distribution of the expression values across samples of an Affymetrix dataset is shown in Fig. 2, specifically for the probeset 201292_at (gene TOP2A). This probeset was highlighted from a previous eBayes analysis, presenting a p-value of 8.02E-34 in this analysis, run to determine differential expression significance. In Fig. 2, a positive correlation between expression and tumor grade is clearly evident.

The platforms to be integrated are from Affymetrix and Agilent so, in this stage, batch effect needs to be considered, as well as ratio-intensity differences between them. Since pre-normalization already done over datasets is not known, it is not possible to establish a concrete mathematical transformation to achieve this goal. It was then chosen to experiment six different ways to obtain a comparison:

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

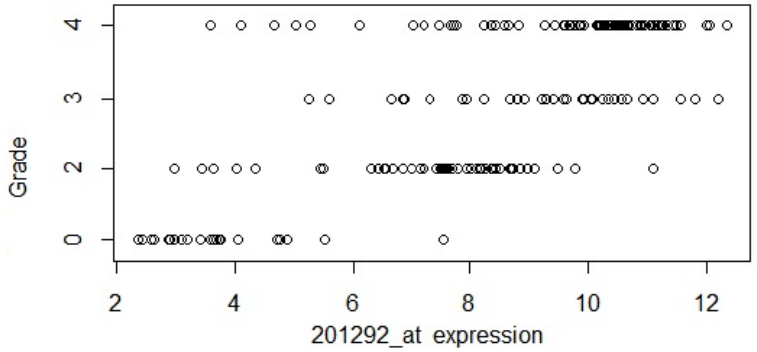

Figure 2: Expression/grade distribution for Affymetrix probeset 201292_at. Grade 0 corresponds to non-tumor samples and grade 1 was not represented on collected datasets

- **Raw.** Where the datasets are joined without any transformation.
- **Norm.** Where the data is mean subtracted and divided by standard deviation.
- **Linear.** A set of reference features are selected, that present the lowest dataset and inter-grade expression variation in all datasets. When joining two datasets, one is used as base and a linear model is built that represents the linear variation between expression values of references from the base and the references of the new dataset. That model is then used to adjust the entire values of the new dataset. This is applied to linear data, without log2 transformation.
- **LinearLog2.** The same as the previous, but applied to log2 transformed data.
- **Grade.** Since the grade is an ordered value, dividing all data by the mean of values of common grade brings the heterogeneous datasets to levels that are comparable. This is applied to linear data, without log2 transformation.
- **GradeLog2.** The same as the previous but applied to log2 transformed data.

4.4 Find candidate biomarkers

Four candidate machine learning algorithms were used, suitable for multi-level classification tasks: Ordinal Logistic Regression (OLR); Support Vector Machines (SVM); k-Nearest Neighbors (k-NN); and Linear Discriminant Analysis (LDA).

The process used to select the best pair of training algorithm - feature set is to firstly choose the possible algorithm candidates and then test each against possible combinations of feature sets. Obviously, this would result on a large processing burden. Then, the first step is to do a filtering stage, to reduce the initial set to a more reasonable subset. Two filtering methods will be used: Pearson's correlation of each feature with the tumor grade and differential expression analysis using empirical Bayes (eBayes), which will give a rank with p-values of the features that best differentiate the four tumor grades. Since some algorithms do not deal well with data collinearity, features that have a high Pearson's correlation among them will be eliminated.

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

With the reduced feature set (rfSet), a greedy wrapper algorithm is deployed to choose, for each algorithm, the best feature subset. The main steps of this algorithm are are shown in Fig. 3.

```
bestFeatureCombinations ← NULL
setSize ← 1
while(setSize <= maxSetSize)
      newCombinations ← allCombinations(bestFeatureCombinations, rfSet)
      model ← trainAlgorithm(newCombinations)
      errorRate ← LOOCV(model)
      bestFeatureCombinations ← bestNewCombinations(errorRate)
      setSize ← setSize + 1
```

Figure 3: Algorithm for best feature set selection

The first iteration is made with only sets with a single feature, testing each algorithm with each of the filtered features. The features which present the best results will be kept and then used as base for the next set of tests, where an additional feature, from the rest of filtered features, is included to organize sets of 2-features combinations. This method is repeated until reaching the *maxSetSize* threshold. The test criterion was the lower classification error rate (number of misclassifications over total number of samples) by Leave One Out Cross Validation (LOOCV). For OLR, the Akaike's Information Criterion (AIC) was calculated, due to lower computation burden.

5 Implementation and Results

Gene expression datasets and phenotypes where collected from Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA), targeting all datasets that present good quality and have samples with explicit brain tumor grade classification. For re-annotation, the probe sequences were obtained from the official sites of the respective vendors and genome annotation was gather from Ensembl. Computational tools were developed with R programming language, using Bioconductor project framework broadly. The database used for datastage and re-annotation processes was implemented over the MySQL Management System.

5.1 Data collection

The data was collected from GEO at the probeset level, while in TCGA the data was found in three levels. This restricted the integration to a level not lower than that of probesets. The Affymetrix and Agilent probeset identifiers were collected from the original datasets. Assuming a specific identifier of a manufacturer measures the same sequence, even if the platform is different, a global set of unique probeset identifiers were collected, for each manufacturer. The Agilent platforms were:

- Human 1A Microarray: Design ID: 011521; Design Format: 1 X 22K;

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

- Human 1B Microarray: Design ID: 011871; Design Format: 1 X 22K;
- Human 1A Microarray (V2): Design ID: 012097; Design Format: 1 X 22K;
- Whole Human Genome Microarray: Design ID: 026652; Design Format: 4x44K v2;
- Whole Human Genome Oligo Microarray: Design ID: 012391; Design Format:1x44K.

The Affymetrix platforms were:

- Human Genome U133 Plus 2.0 Array
- Human Genome U133A
- Human Genome U133B
- Human Genome U95Av2

The Agilent probe sequences were collected with the eArray tool, available in the following URL: https://earray.chem.agilent.com/earray, using the option Microarray → Browse Microarray Designs, selecting "H. sapiens". The Affymetrix sequences were taken from the Web site through the option Products → Microarray Solutions → 3' IVT Expression Analysis → select desired microarray → Technical Documentation → Sequence Files (Tabular). It was checked that all probesets of U133A and U133B are also in U133 Plus 2.0 and that some U95Av2 probesets are also in U133 Plus 2.0. So, the collection was started with the 54613 U133 Plus 2.0 probesets, to which 12286 U95Av2 specific probesets were added, making a total of 66899 unique probeset identifiers. On U133 Plus 2.0 array, 9711 probes were identified with duplicate sequences and in the U95Av2 array 1722 duplicate sequences were found. After joining the two sets, repeated sequences were noted in 22351 probes.

5.2 Re-annotation

Using Biostrings, with the human genome version BSgenome.Hsapiens.UCSC.hg19, the match of each probe sequence with the genome was done. Each probe could have zero or more matches along the genome. A maximum of one base pair (bp) mismatch [11] was allowed. From the 799238 Affymetrix probes less than 5% do not have any matches in the genome. Most probes present only one match, while a maximum of 366059 matches was registered for a single probe. The largest numbers of matches were recorded in 4 probes of the probeset 1008_f_at. There are also cases where the same probe matches in several distinct regions of the same gene. For example, in the probeset 1557055_s_at a probe was detected with 10 distinct matches in the gene RP11-206L10.11 (ENSG00000228794). The distribution of number of matches per probe is given in Tab. 2.

For some probesets (0,6%), no match has been identified for any probe. This could be explained by the upgrade of the genome annotation. Other possibilities could be pointed, like the fact that approximately 200 Mbp of theTab human genome, mainly from the centromeres and

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Table 2: Number of genome matches per probe

matches	1	2	3	4	5
probes	675561	44974	13528	6229	4097
matches	6	7	8	9	10
probes	2776	1761	1502	1106	835

the short arms of the acrocentric chromosomes, are missing from the human reference genome [26]. The Genome Reference Consortium (GRC) builds the genome reference leaving these sequences in structures outside chromosomes. A GRC genome assembly, for H. Sapiens species, is composed by [27]: 24 "relatively complete" sequences for chromosomes 1 to 22, X and Y; a complete mitochondrial sequence; several "unlocalized sequences" (their exact location within the chromosome is not known); several "unplaced sequences" (their chromosomal association is not known); and several "alternate loci" (contain alternate representations of specific regions). Considering these 'extra' genome sections, 487 additional probes were identified. The rest of the probes were analyzed with BLAST tools from Ensembl and UCSC, but continued to return no matches.

Annotation through GeneAnnot [28] presented incomplete results (Table 3) and is gene centric, i.e., it is not possible to obtain the exact chromosome location. The information obtained from GeneAnnot can be found through Affymetrix annotation files. The fact that some probes do not match any region of the genome could be due to outdated Affymetrix probe design (Build 133, April 2001 for U133 Plus 2.0 Array) when compared with the genome assemblies that are frequently updated by GRC (February 2009 for GRCh37/hg19). Probes that showed no match were discarded.

Table 3: Some Probesets Annotation by GeneAnnot

Probeset	**Gene**	**Chromo.**	**Strand**	**Start**	**End**
224344_at	COX6A1	12	+	120875893	120878545
1565535_x_at	M74301	?	?	?	?
1554232_a_at	BC018433	?	?	?	?
208303_s_at	CRLF2	X	-	1314869	1331616

Among the probes with matches, 95.2% registered at least one match in the genes referenced by Ensembl, getting 5% of the probes without apparent connection to any gene. 79.6% of the probes with matches in genes had at least one match in an exon. The remaining cover zones of introns, UTRs, or boundary regions that do not belong in their totality to exons. The annotation process was made by stages, crossing the probe matches obtained from Biostrings with the genome information contained on the Ensembl database. From Ensembl, for each exon, information was collected about the corresponding gene, sequence, location (start and end) on the respective chromosome, and strand. A filter was applied to return only the exons present in the 24 chromosomes from 1 to Y. The information extraction for exons was made using biomaRt R package, because it allows to associate the sequence of each exon to the rest of the information. The transcript annotation data obtained with the biomaRt package include non-normalized fields that would require a considerable processing time for their normalization. Thus, the Ensembl data for transcripts and genes was collected using the BioMart web tool in CSV format from http://www.ensembl.org/biomart/martview, with the options: Database: Ensembl Genes 72 $\rightarrow$ Dataset: Homo sapiens genes (GRCh37.p11) $\rightarrow$ Attributes. It was possible to identify,

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

for each transcript, the gene, the strand, start and end positions on the chromosome. In parallel, the transcript-exons map was built.

5.3 Datasets integration

Three datasets came from GEO, built with Affymetrix HG-U133Plus2.0, HG-U133A and HG-U133B platforms: GDS1962 (23 samples grade 0, 45 g2, 31 g3, 81 g4); GDS1815_6 (24 g3, 76 g4); and GDS1975_6 (26 g3, 59 g4). Two datasets were taken from TCGA, built with Agilent platforms: TCGA_AGIL2 (10 g0, 232 g4); and TCGA_AGIL4 (7 g2, 20 g3). As GDS1962 is the most complete and balanced dataset, it was used for preliminary Affymetrix probesets tests. For Agilent, TCGA_AGIL2 g4 subset was trimmed, leaving only 40 g4 samples. The two Agilent datasets were joined on TCGA_AGIL2_4 (10 g0, 7 g2, 20 g3, 40 g4). Higher-level datasets were built from probeset data. Gene level datasets were produced averaging all probesets associated with each gene. Transcript datasets were produced by weighted average of the probesets associated with each transcript (Eq. 5) [29].

5.4 Find candidate biomarkers

Pearson's correlation and eBayes filters were applied on these datasets and the probesets were sorted by rank. Over this, a cleaning process was performed to remove collinear probesets, using a 0.9 Pearson's correlation cutoff. Finally, only a set of the 200 best ranked probesets were kept for further tests. The wrapper method was tuned to make combinations until a maximum of 20 features. Also, the quantity of the best combinations preserved between cycle iteration was limited to 20. For k-NN, k was always set to 3. The main results are provided in Tables 4, 5 and 6.

Table 4: Results for classifiers (columns), filtering methods (rows) at the probeset level. #P - number of probesets; ER - Error Rate

		OLR	SVM	k-NN	LDA
Pearson Cor.	Affy	#P: 16 ER: 0.25	#P: 13 ER: 0.11	#P: 18 ER: 0.14	#P: 13 ER: 0.19
	Agil	#P: 02 ER: 0.13	#P: 6 ER: 0.04	#P: 06 ER: 0.01	#P: 05 ER: 0.04
eBayes	Affy	#P: 16 ER: 0.28	#P: 10 ER: 0.09	#P: 20 ER: 0.14	#P: 12 ER: 0.17
	Agil	#P: 05 ER: 0.17	#P: 09 ER: 0.03	#P: 06 ER: 0	#P: 04 ER: 0.03

Table 5: Results for classifiers (columns), filtering methods (rows) at the gene level. #G - number of genes; ER - Error Rate

		OLR	SVM	k-NN	LDA
Pearson Cor.	Affy	#G: 16 ER: 0.26	#G: 19 ER: 0.08	#G: 14 ER: 0.14	#G: 19 ER: 0.17
	Agil	#G: 03 ER: 0.12	#G: 08 ER: 0.03	#G: 05 ER: 0.01	#G: 11 ER: 0
eBayes	Affy	#G: 16 ER: 0.23	#G: 17 ER: 0.09	#G: 18 ER: 0.14	#G: 19 ER: 0.17
	Agil	#G: 05 ER: 0.17	#G: 04 ER: 0.01	#G: 06 ER: 0	#G: 13 ER: 0

The differences between results associated with the filtering method are not significant, so one will be used arbitrarily. It is also possible to see that the accuracy is always better for Agilent datasets, and SVMs have the best performance when considered both platforms. In the matters

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Table 6: Results for classifiers (columns), filtering methods (rows) at the transcript level. #T - number of transcripts; ER - Error Rate

		OLR	SVM	k-NN	LDA
Pearson Cor.	Affy	#T: 18 ER: 0.25	#T: 14 ER: 0.1	#T: 20 ER: 0.14	#T: 17 ER: 0.17
	Agil	#T: 03 ER: 0.09	#T: 06 ER: 0.03	#T: 07 ER: 0.01	#T: 17 ER: 0.04
eBayes	Affy	#T: 18 ER: 0.23	#T: 18 ER: 0.07	#T: 17 ER: 0.13	#T: 13 ER: 0.16
	Agil	not converge	#T: 06 ER: 0.02	#T: 04 ER: 0.02	#T: 06 ER: 0

of probeset, gene or transcript level, the former presented slightly worse results. Between gene and transcript, it is not possible to distinguish a clear difference. Besides the identification of the best classification algorithm, it is important to compare the prediction features obtained for Affymetrix and Agilent. If a transcript/gene is related to the grade phenotype, then it should be relevant for both platforms. To check this premise, eBayes was applied to rank features separately, and then pairwise Kendall's coefficient of concordance (W) was calculated. For genes, W = 0.675 was obtained, and for transcripts W = 0.711. There is a considerable discordance about feature relevance between Affymetrix and Agilent. Deepening the analysis, the features selected as the best predictor with SVM for Affymetrix was used for Agilent tests and vice versa. With Affymetrix predictors, a LOOCV error of 0.10 for genes and 0.09 for transcripts, was obtained when these predictors were applied on Agilent. For Agilent predictors, an error of 0.29 for genes and 0.32 for transcripts was obtained, when applied on Affymetrix data. It is clear that predictors obtained from Agilent data tend to overfit more and loose generalization.

The next tests were made joining all datasets on a single one, using different methods for data integration (Table 7). SVMs are confirmed as the training algorithm that gives the best results and the division of the data by a reference set of values (grade) also improves accuracy. As the common grade between all datasets is g3, this was the reference grade used. Joining heterogeneous datasets aims to produce a robust training set that allows to classify a new sample with a good rate of assertiveness. With these LOOCV tests it is possible to preliminary confirm that it is possible to do this with datasets from Agilent and Affymetrix.

Table 7: LOOCV error rate for Affymetrix and Agilent integrated data

		Raw	Norm	Linear	LinearLog2	Grade	GradeLog2
OLR	Genes	0.24	0.28	0.27	0.31	0.2	0.2
	Transcripts	0.24	0.3	0.3	0.26	0.19	0.23
SVM	Genes	0.11	0.09	0.11	0.09	0.08	0.12
	Transcripts	0.09	0.1	0.12	0.15	0.08	0.11
k-NN	Genes	0.13	0.16	0.19	0.15	0.14	0.17
	Transcripts	0.11	0.18	0.19	0.2	0.14	0.17
LDA	Genes	0.16	0.2	0.19	0.16	0.14	0.18
	Transcripts	0.16	0.19	0.22	0.19	0.15	0.18

To take the tests a little further, the same joined dataset was used to build a model to predict a new Affymetrix dataset, collected from TCGA, not used before in any stage. This new dataset has 10 g0 and 277 g4 samples so, to apply the reference grade method, all data was divided by grade 0. The results again confirmed the good performance of the method (Table 8).

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Table 8: Confusion matrices for prediction of the new dataset

Genes		Predicted			
	Grade	0	2	3	4
Observed	0	10	0	0	0
	4	0	6	5	266
Error rate:	0.038				

Transcripts		Predicted			
	Grade	0	2	3	4
Observed	0	10	0	0	0
	4	0	4	8	265
Error rate:	0.042				

6 Conclusions

It is rare to find a work that integrates data from platforms with a different measurement basis, like the log intensities and log ratios that are addressed in this work. The different feature annotation is also an obstacle for this kind of integration, which was overpassed by re-annotation at transcript level, a crucial step in the integration of gene expression data from different vendors. The original probeset-oriented annotation makes impossible the direct integration of data due to the use of different identifiers. Since each manufacturer decides which regions of the genome should be under measurement, and this is not a static assignment, a flexible method of re-annotation is a great achievement. Gene-level integration can generate divergent results, given alternative splicing. Confining expression information to a gene level basis means losing precious information about the different biological functions of genes. The separate analysis of microarray data followed by result integration is a common alternative, but suffers from the problem of few samples per experiment, which can lead to skewed conclusions. The option for a transcript-oriented integration, presented in this work, preserves the proteomic structure, allowing a biological meaningful integration between data sets of different vendors. This approach improves the strength of candidate features, by allowing to exclude those with divergent behavior between platforms, and associating each one to a specific function.

In this work, the process of choosing the best brain tumor grade predictors set and the best classification algorithm was addressed. The filter methods reduce drastically the initial dimension of the datasets, and did not present great differences among them. SVMs stood out as the best classifiers for this case. The use of a reference grade to bring data to equivalent scales between datasets also shown to be a good method of integration. The level of data to use on the integration, gene or transcript, needs a more profound interpretation. The probesets used for expression measurement are exon centric and do not read all exons of a gene. Since the same exon could be part of several different transcripts, it is common to find many of them with the same expression value, which causes entropy to learning algorithms. Even filtering the 'clones', in many cases the expression value of a transcript is the same as the correspondent gene, when all the probesets measuring the gene also measure that transcript. It is than clear that it is necessary to upgrade the microarray expression measurement to be possible to use transcript level effectively. The final model has shown good results when applied to a new dataset, showing that it is possible to achieve a greater level of generalization based on heterogeneous microarray dataset integration.

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Acknowledgements

The authors thank the FCT Strategic Project of UID/BIO/04469/2013 unit, the project RECI/BBB-EBI/0179/2012 (FCOMP-01-0124-FEDER-027462) and the project BioInd - Biotechnology and Bioengineering for improved Industrial and Agro-Food processes", REF. NORTE-07-0124-FEDER-000028 Co-funded by the Programa Operacional Regional do Norte (ON.2 O Novo Norte), QREN, FEDER.

References

[1] J. Seznec and U. Naumann. Microarray analysis in a cell death resistant glioma cell line to identify signaling pathways and novel genes controlling resistance and malignancy. *Cancers*, 3(3):2827–2843, 2011.

[2] A. Perez-Diez, A. Morgun and N. Shulzhenko. *Microarrays for Cancer Diagnosis and Classification*, volume 593 of *Advances in Experimental Medicine and Biology*. Springer New York, 2007.

[3] P. S. Mischel, T. F. Cloughesy and S. F. Nelson. Dna-microarray analysis of brain cancer: molecular classification for therapy. *Nature Reviews Neuroscience*, 5(1):782–792, 2004.

[4] S. B. Cho and J. Ryu. Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. In *Proc. of the IEEE*, volume 90, pages 1744–1753. 2002.

[5] A. Lorena, I. Costa and M. de Souto. On the complexity of gene expression classification data sets. In *Hybrid Intelligent Systems, 2008. HIS '08. Eighth International Conference on*, pages 825–830. 2008.

[6] J. Huang, H. Fang and X. Fan. Decision forest for classification of gene expression data. *Computers in Biology and Medicine*, 40(8):698–704, 2010.

[7] C. Stretch, S. Khan, N. Asgarian et al. Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature. *PLoS ONE*, 8(6), 2013.

[8] D. Elizondo, B. Passow, R. Birkenhead and A. Huemer. *Principal Manifolds for Data Visualization and Dimension Reduction*, volume 58 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag Berlin Heidelberg, 2008.

[9] J. K. Choi, U. Yu, S. Kim and O. J. Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(suppl 1):i84–i90, 2003.

[10] W. E. Johnson, C. Li and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[11] B. Ballester, N. Johnson, G. Proctor and P. Flicek. Consistent annotation of gene expression arrays. *BMC Genomics*, 11(1):294–307, 2010.

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

[12] E. Rouchka, A. Phatak and A. Singh. Effect of single nucleotide polymorphisms on affymetrix match-mismatch probe pairs. *Bioinformation*, 2(9):405–411, 2008.

[13] Affymetrix. Statistical algorithms description document. Technical Report 701137 Rev 3, 2002.

[14] M. Zahurak, G. Parmigiani, W. Yu, R. Scharpf, D. Berman, E. Schaeffer, S. Shabbeer and L. Cope. Pre-processing agilent microarray data. *BMC Bioinformatics*, 8(1):142–154, 2007.

[15] J. H. Do and D.-K. Choi. Normalization of microarray data: Single-labeled and dual-labeled arrays. *Molecules and Cells*, 22(3):254–261, 2006.

[16] D. Stekel. *Microarray Bioinformatics*. 1. Cambridge University Press, 2003.

[17] H. Yu, F. Wang, K. Tu, L. Xie, Y.-Y. Li and Y.-X. Li. Transcript-level annotation of affymetrix probesets improves the interpretation of gene expression data. *BMC Bioinformatics*, 8(1):194–208, 2007.

[18] A. Culhane, G. Perriere and D. Higgins. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4(1):59–73, 2003.

[19] Y. Woo, J. Affourtit, S. Daigle, A. Viale, K. Johnson, J. Naggert and G. Churchill. A comparison of cdna, oligonucleotide, and affymetrix genechip gene expression microarray platforms. *Journal of Biomolecular Techniques*, 15(4):276–284, 2004.

[20] A. K. Jarvinen, S. Hautaniemi, H. Edgren, P. Auvinen, J. Saarela, O. P. Kallioniemi and O. Monni. Are data from different gene expression microarray platforms comparable? *Genomics*, 83(6):1164–1168, 2004.

[21] M. Bellis. Mapping of affymetrix probe sets to groups of transcripts using transcriptional networks. *ARXIV*, eprint arXiv:1201.2033, 2012.

[22] E. Kostadinova. Data integration: An approach to improve the preprocessing and analysis of gene expression data. *Union of Scientists in Bulgaria*, 6(1):120–133, 2013.

[23] C. Meng, B. Kuster, A. Culhane and A. Gholami. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, 15(1):162–174, 2014.

[24] A. Papiez, P. Finnon, C. Badie, S. Bouffler and J. Polanska. Integrating expression data from different microarray platforms in search of biomarkers of radiosensitivity. In *In proceedings of IWBBIO*, volume 2, pages 484–493. 2014.

[25] C. D. Willis, K. Lafferty-Whyte, M. Baker and R. G. Pestell. Integrating transcriptomic data using metacore pathway analysis to identify novel biomarkers of bevacizumab target engagement. *Cancer Research*, 74(19):352–359, 2014.

[26] G. Genovese, R. Handsaker, H. Li, E. Kenny and S. McCarroll. Mapping the human reference genomes missing sequence by three-way admixture in latino genomes. *The American Journal of Human Genetics*, 93(3):411–421, 2013.

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

[27] G. Asimenos. Human genome, 2014. URL https://wiki.dnanexus.com/Scientific-Notes/human-genome.

[28] F. Ferrari, S. Bortoluzzi, A. Coppe, A. Sirota, M. Safran, M. Shmoish, S. Ferrari, D. Lancet, G. Danieli and S. Bicciato. Novel definition files for human genechips based on geneannot. *BMC Bioinformatics*, 8(1):446–451, 2007.

[29] E. Valente and M. Rocha. Transcript-based reannotation for microarray probesets. In *Proc. 30th ACM/SIGAPP Symposium On Applied Computing (SAC2015)*. 2015.

Copyright 2015 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).