Table of contents:

Summary 1
Introduction 3
Description of Functionality 7
Outline of Backend Technologies 9
Using the System 10
References 19
Publications arising from this study 19
Summary

DNA microarrays are widely used for gene expression profiling. Raw data resulting from microarray experiments, however, tends to be very noisy and there are many sources of technical variation and bias. Therefore, raw microarray data needs to be quality assessed and interactively preprocessed to minimise variation before statistical analysis in order to achieve meaningful results.

Bioconductor is an existing open source software project that attempts to facilitate analysis of genomic data. It is a collection of packages for the statistical programming language R. Bioconductor is particularly useful in analyzing microarray experiments. The problem is that the R programming language's command line interface is intimidating to many users who do not have a strong background in computing. This often leads to the use of commercial software which is expensive and less effective statistical techniques.

This project aims to bridge this gap by providing a user friendly web-based interface to the cutting edge statistical techniques of Bioconductor. BioconductorBuntu is a custom distribution of Ubuntu Linux that wraps the analysis tools developed by the project in an easily installable and distributable format. The server is setup by running a straightforward installation CD.

Analysis tools that we have constructed to date complement existing analysis of microarray data. However, there is need for further development of the analysis pipeline to address current needs and future development.

The initial scope of this project primarily focused on analysis of Affymetrix GeneChip arrays. However, the facilities for basic analysis of dual dye cDNA arrays and single dye Exiqon miRNA arrays have also been implemented and provide a solid foundation for future development.

BioconductorBuntu is a custom distribution of Ubuntu Linux that wraps the analysis tools developed by the project in an easily installable and distributable format. The server is setup by running a straightforward installation CD.

The system is best installed on a dedicated server, allowing any individuals connected to the same network to make use of the analysis tools hosted on the server.
Introduction

Eukaryotes are organisms whose cells are organized into complex structures enclosed within membranes. Most eukaryotic organisms, for example, human beings, contain billions of individual cells. Almost all of these cells contain, within each nucleus, the entire genome for that organism. This genome contains the organism's complete hereditary information in the form of deoxyribonucleic acid (DNA), that encodes a complete blueprint for all activities and attributes of the organism.

In the human body, the genome consists of 23 pairs of chromosomes. One of each of these pairs is inherited from the mother and the other from the father. Each chromosome consists of chains of DNA, consisting of two polymers made up of units called nucleotides. Each nucleotide consists of a deoxyribose sugar, a phosphate group and one of the four nitrogen bases, guanine, adenine, thymine and cytosine. These bases, which are usually represented by their first letters, G, A, T and C, are where hereditary genetic information is actually encoded. It is worth noting that one of the two strands of the DNA double helix will suffice to describe this information; this is because of complementary base pairing, whereby an A on one strand always binds to a T on the other and a C always binds to a G.

Genes are essentially segments of the DNA structure described above. Loosely speaking, a gene is a section of DNA that defines a single trait by encoding a particular pattern, about 27,000 of which exist in humans. Often though, we are faced with a case where protein-coding sequences have no clear beginning or end; more technically, a gene is a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcriptional control regions, also known as introns and/or other functional sequence regions.

The main purpose of genes is to serve as blueprints for the creation of proteins. Proteins are made of amino acids and are responsible for the structure and activity of an organism at a cellular level. They can control cell division, protein synthesis, and allow for cellular recognition. This process is performed by so-called DNA polymerases.
After transcription, the mRNA molecule leaves the nucleus of the cell where it is translated into a protein in a process called translation. This process involves the ribosomes, which read the code carried by mRNA molecules from the cell nucleus and create proteins by combining any of the 20 amino acids into complex polypeptide chains. These proteins are the building blocks of the organism. This process of translating a gene into a functional product is known as gene expression.

DNA microarrays are a high-throughput technology used to measure the expression levels of thousands of genes in most cases all of the genes in a genome simultaneously. The fundamental idea behind most microarrays is to exploit complementary base pairing (see previous section) to measure the amount of the different types of mRNA molecules in a cell, thereby indirectly measuring the expression levels of the genes that are responsible for the expression of those particular mRNA molecules.

The gene on a microarray consists of a single stranded DNA oligonucleotide—called a probe. Each probe is designed to bind to a specific mRNA molecule. The probe is typically a synthetic DNA molecule that includes a complementary sequence to the gene that it targets. The mRNA molecules are extracted from the original cell and have previously been labelled with a fluorescent dye. This allows the amount of hybridisation to be measured by the level of fluorescence produced by the dye which is examined with a scanner. The scanner then outputs a text file for each array, which contains the relevant data pertaining to that array, such as the level of fluorescence of each spot and the level of background noise. These text files are subsequently computationally analysed. In theory, a spot with brighter fluorescence means that more mRNA has hybridised to that probe in the original cell and hence the gene represented by that spot is experiencing a higher level of expression.

The type of DNA microarray most widely used today can be broadly divided into two categories: cDNA arrays and oligonucleotide arrays. The GeneChip, which is manufactured by Affymetrix, is an oligonucleotide array and is the most commonly used type of microarray today. Each array will contain hundreds of thousands of probes, each of which is a mixture of thousands of probes representing all of the genes that are transcribed in that cell. The GeneChip scanner can read the amount of fluorescence emitted by each probe and hence the expression level of the gene it represents.
Each gene that is being targeted is represented by typically (but not necessarily always) 11 pairs of these probes. This set of probes contains 11 perfect match (PM) probes, which are exactly complementary to the DNA sequence of a subset of 25 bases of the target gene. Each PM probe has a corresponding mismatch probe (MM), which contains the same 25 base long sequence as the PM probe, except for the fact that the middle base, or the 13th base in the chain, is substituted for the complement of the 13th base of its corresponding PM probe; so for example, a G in the 13th base of a PM probe will be replaced with a C in the MM probe. This is meant to get an estimate of non-specific binding, which occurs when mRNA from a gene is not targeted.

Figure 1. Affymetrix Genechip. Perfect match (PM) and mismatch (MM) probe pairs make up the set of probes representing a gene transcript. The middle base of the MM probe differs from the PM probe.
The cDNA molecules are usually more than 500 bases long. Each of these samples is generally a control sample, while the other is a sample of interest such as a tumour sample. Each of these samples is labelled with a particular dye: red (Cy5) or green (Cy3). When the array is read, the measure of how much cDNA binds to its corresponding spot gives an accurate measure of the expression level of the gene in question.

Instead of expression levels of an individual sample being measured directly, two separate samples are hybridised to the same array at the one time. One of these samples is a control sample while the other is a sample of interest such as a tumour sample. Each of these samples is labelled with a particular dye: red (Cy5) or green (Cy3). When the array is read, the measure of how much cDNA binds to its corresponding spot gives an accurate measure of the expression level of the gene in question.

In addition to measuring expression levels of individual samples, cDNA microarrays are also used to detect differences in expression between two samples. Two separate samples are hybridised to the same array at the one time. One of these samples is a control sample, while the other is a sample of interest such as tumour tissue. Each of these samples is labelled with a particular dye: red (Cy5) or green (Cy3). When the array is read, the measure of how much cDNA binds to its corresponding spot gives an accurate measure of the expression level of the gene in question.

Microarrays can also be used to measure expression levels of miRNAs. miRNAs are short RNA molecules, generally about 22 nucleotides in length. They are encoded in genes but are not translated into proteins. Instead, these molecules downregulate the expression of certain genes. They achieve this by being complementary to specific mRNA molecules created in a cell. The miRNA molecules bind to the complementary sections of these mRNAs, preventing them from being translated into proteins.

Frogmin is a naturally occurring miRNA that is found in frogs. It is known that this miRNA binds to a specific mRNA that is expressed in the skin of frogs. This binding prevents the expression of a specific gene that is associated with the skin of frogs. This binding prevents the expression of a specific gene that is associated with the skin of frogs. In other words, the presence of this miRNA prevents the expression of this particular gene in frogs.
hybridising the sample to the array and reading the hybridisation levels with a scanner are similar to those of other arrays.

Description of Functionality

The web based tool described in the following sections can be accessed either on http://europa.it.nuigalway.ie/cgi-bin/login.py or http://10.12.1.45/cgi-bin/login.py and these sections are best read in conjunction with the online tool.

Following connection to the server via a web browser, each user must create an account on the system via the 'Register' link, uploading their relevant details. After registration, the user may thereafter access the system and upload raw data. Uploaded data are assumed to be in a '.zip' archive, whose data files are unpacked and stored server-side prior to analysis. Currently, a user has three options, to upload Affymetrix R, dual dye or single dye data. For each of Affymetrix R experiments, the user has the option of uploading full-length raw data or a compendium CDF file instead of the default Bioconductor CDF file. At upload, the user can select Affymetrix R data to be mapped to previously created Affyprobeminer's remapped CDF files. An entry containing all relevant information is created in the database for the new experiment as data are being uploaded and processed, progress is printed on the screen, allowing users to monitor this process in real-time. After data upload, the user is presented with a screen that requires them to specify the number of factors for the experiment and the number of levels for each of these factors, as well as names for the levels and factors. Next they must assign the appropriate levels to each factor to each array.

The next stage is quality control, to ensure that data are of sufficient integrity. There are a number of options available here such as boxplots, histograms, PCA and many more diagnostic plots on normalised and unnormalised data. Several of the most commonly used normalisation options are also available. The user has the option of selecting the appropriate options on the upload screen or choosing them after data upload. Based on the output of quality control, it is up to the user to decide if some or all of the arrays need to be redone or removed from differential expression analysis. The 'Data Select' link allows the removal of arrays. These decisions are then reflected in the analysis.
The next and most significant step is differential expression analysis. Using Affymetrix R data, the user has the option of using the Limma or PUMA packages to perform this analysis. Under limma a number of preprocessing methods, such as RMA, GCRMA, MAS5 and custom methods, specifying the steps taken for normalisation, background correction, PM correction and summarisation, are available. If the user opts to use the PUMA package, data are preprocessed using PUMA, which implements global median scaling normalisation. The expression level information from the different arrays of each condition is then combined. Differential expression is then calculated using the Probability of Positive Log Ratio (PPLR) method. Because PPLR analysis can be time-consuming, users are notified by email when the analysis is complete. For dual dye analysis the same normalisation methods that are available in the Limma package. There are a number of different options for normalisation between arrays, normalisation within array, and background correction. All of these options are available in both Limma and PUMA packages. After the analysis is completed, a ranked gene list, containing such metrics as fold change, p-values, adjusted p-values and B-statistic, is output to the viewing screen. This list contains links to gene annotation information which is retrieved using the Bioconductor package BioMart and is dynamically loaded using AJAX. The ranked gene list may also be downloaded as an Excel file. All of these steps are available in both Limma and PUMA packages. The user also has the option to download the .Rdata file containing the complete R workspace environment for a particular analysis. This may be useful to continue analyses from the command line, which can be more flexible in certain situations and useful to advanced users. The user also has the option to delete previous experiments and analyses. Users may also manage their account by defining new experiments and analyses.
The project is entirely based on open source software which is free to redistribute, use and alter. Ubuntu Linux is the operating system upon which the server is built. Ubuntu was chosen for several reasons, one of which is that it is the most commonly used Linux distribution, meaning it is likely to be familiar to a wider user base as well as being a stable operating system. Ubuntu is often used in large organisations for its security and stability. The distribution was customised using an open source script called remastersys, which allows its users to make changes to a basic Ubuntu installation and save them as a remastered installation CD. All analysis of gene expression microarray data are performed in R, specifically using Bioconductor. This distribution is available to download as Bioconductor Buntu from http://bioinf.nuigalway.ie. Pipelines facilitating the analysis of other microarray platforms are also possible to the user community by reconfiguring Bioconductor Buntu accordingly. The CGI scripts which run on the server are written in Python. These scripts handle input and output to and from the R and MySQL database, as well as making calls to the database and writing data to it. The Apache web server is used and the powerful open source database MySQL is used to store and retrieve various forms of data. Exim is used as a mail transfer agent and Devcot as an IMAP server. Other technologies used by the server include Imagemagick for image manipulation and particularly creating thumbnail versions of image files and PhpMyAdmin for web based database administration. Considering the large data footprints associated with typical microarray analyses, Bioconductor Buntu should ideally be installed on a high end server; centralising the processing of such large amounts of data on a high specification machine particularly for complex analysis will lead to greatly improved performance compared to analysis on a standard desktop machine.
Using the System

The first thing a user must do is register with the system. This is achieved by clicking the "Register" tab on the top of the page. The "register.py" script is invoked on clicking this tab and handles the process. This page contains several JavaScript functions to authenticate that the form has been filled out correctly, ensuring that the email address supplied is valid and all required fields have been completed.

If the form is not properly filled out or contains a duplicate entry, the user will receive an error notification and be prompted to fill out the form correctly. The user must then proceed to login for the first time. The first stage in creating an experiment is to upload the raw data. At this stage, the user is presented with a page that displays options to upload different types of raw data (Figure 2).

![Figure 2: Screen shot showing uploading of data to Microarray analysis system](image)

Figure 2. Screen shot showing uploading of data to Microarray analysis system.
For all data types the user must package their raw data files in a zip archive. This can be accomplished with any number of freely available compression tools like 7-Zip or WinZip. Zipping the data has the double advantage of significantly compressing the data for faster upload and also means that only one file needs to be uploaded. The user must also specify an experiment name every time they upload a new dataset. There are some differences in the rest of the information supplied when uploading different types of experiments. For Affymetrix data, the user has the choice of using the default Affymetrix supplied CDF file or using the remapped CDF file from AffyProbeMiner. For Dual dye data, the user may wish to specify a "Spot Types" file. This file contains information which describes the functional groups on the array and can be used to identify sample control signals. Depending on the type of data being uploaded, a GenePix Array List (GAL) file may also be required. This file contains information regarding the physical layout of the array used in the experiment and is needed to generate array images, as well as to determine the amount of data to be analyzed and the output options. Our bovine dataset, which includes 12 arrays, weighs in at a hefty 41 megabytes, even when compressed. This upload will take several minutes, even on a fast connection.
The first screenshot on the next page shows the screen directly after the upload button has been clicked. This is what is seen while the actual data transfer is in progress. Note how the user is informed to be patient and told that the upload will take several minutes while the spinning animation gives the impression that something is happening. The next screen shows what has happened after data is uploaded. For this dataset, this screen takes about 30 seconds to work to completion. Crucially, progress is printed as each step is completed, so the user is not presented with a static screen until the page is fully loaded.

The next stage of almost any analysis is to assign the experiment’s phenotypic data. This stage differs significantly for Affymetrix datasets, where the user has the option to specify an experimental design of up to 10 factors, with up to 10 levels of each factor. To clarify what I mean by this, consider our bovine dataset from earlier. The simplest experiment contains only one factor, our bovine experiment for example contains one factor, which we could call “Negative Energy Balance”; this factor is then said to have two levels, one of which is the negative energy balance group and the other of which is a control group. An appropriate level of each factor is then assigned to each array to describe the experimental design. For example, an array might be assigned to the negative energy balance group or to the control group. The next stage is to describe the phenotypic state of the sample. In the above example, arrays will either be designated as being from the negative energy balance group or from the control group, based on which of either level of the single factor they are assigned.

We will now consider the more complex experimental design of the Estrogen Dataset. This is an example of an experiment that is handled with Bioconductor. It is a 2x2 factorial design experiment on Affymetrix HGU95av2 arrays. The aim of the study is to identify genes which respond to estrogen and to classify these into early and late responders. The experiment is set up using the popular 2x2 factorial design. It contains two factors, both of which have two levels. The first factor defines the length of exposure of the samples, either 10 or 48 hours, and is said to have the levels “+10” and “+48”. The second factor defines the estrogen treatment, which is either present or absent, and is said to have the levels “Present” and “Absent”. Assignment of the phenotype data to each array is done by defining these factors in the experimental design editor, and then selecting the levels for each factor. This allows us to describe the two-way interaction between factors, and to perform differential expression analysis.
down menus that define the amount of factors or levels is changed, more text boxes are dynamically added without reloading the page. This is achieved using JavaScript.

The next step in any analysis is to run quality control. The initial screen displays a list of available quality control options that can be selected using checkboxes. These options can be used to assess raw, preprocessed data, or both.

If working with dual dye or Affymetrix data, hovering the mouse pointer over the "Preprocessed Data" heading on the table allows the user to select from a number of different preprocessing methods. Changing this value will change how data is preprocessed for the appropriate plots.

Hovering the mouse pointer over any particular quality control option will cause a tooltip to appear that gives an outline of how that particular plot or metric can be used and how it should be interpreted. This is a useful feature for novice users.

Once the required options have been selected, the user will submit the form. The next page displays everything that has been requested. Quality control plots are displayed as thumbnail images that can be viewed by clicking the thumbnail. Additionally, hovering the mouse pointer over any particular plot or metric will cause a tooltip to appear that gives an outline of how that particular plot or metric can be used and how it should be interpreted. This is a useful feature for novice users.

It is again important to note that the output of this page is piped directly to the screen as the page is generated. If every option is checked, the whole page takes about a minute to print to completion for our bovine dataset, but the user is never left staring at a static or blank screen. Even the progress of any preprocessing method that is being undertaken is piped directly from R to the user's screen.

Additionally, if analysing an Affymetrix experiment and the user selects the "PUMA PCA and Scree Plot" option, instead of these plots being generated there and then, the user will instead be informed that notification of completion of the plots will be emailed to them.
Due to the higher volume of data the PUMA method deals with, the visualization stage may take a longer time to complete, approximately 40 minutes. After completing the analysis using the Quality Control tool under the Saved Analysis menu, the data selection phase may be required following quality control. In certain situations, users may decide that an array is not suitable for inclusion in further analysis. The data selection page allows users to exclude an array from subsequent analysis without having to create a new experiment. For example, if we were to exclude the array "NS7.CEL" from our bovine dataset for differential expression analysis, we can do so by simply unchecking the array in the page.
The user begins an analysis by choosing a differentially expressed gene set from a pre-defined list, and then proceeds to the differential expression analysis. There are numerous different gene sets available at this stage and options differ significantly based on the type of dataset involved.

In the case of Affymetrix data, the user selects between the Limma and PUMA packages and chooses a preprocessing method. From this point onward, the user can select a combination of options that will be used to identify differentially expressed genes. This is repeated for each dataset, as there is no requirement to have differentially expressed genes in the same dataset.

The next step in an analysis is to run differential expression. There are numerous different choices available at this stage and options differ significantly based on the type of dataset involved.

In the case of Affymetrix the user selects between the Limma and PUMA packages and selects a preprocessing method, after which they are required to decide on which contrasts they wish to test for differential expression. This is simple for our bovine dataset, as it is an experimental design with only two levels and a single factor. In such a situation there are obviously several contrasts available, some of which may be worth examining and some which may not. Two contrasts which are obviously of interest are "Present +10 VS Absent +10" and "Present +48 VS Absent +48" which will tell us which genes are calculated as being differentially expressed between the estrogen absent and estrogen present groups at 10 hours and subsequently at 48 hours. These results can be downloaded as a .xls format spreadsheet file, which can be viewed in an application such as Microsoft Excel or OpenOffice.

If the user opts to use the PUMA method to assess differential expression, they must again specify the contrasts of interest; once these are submitted a message is printed notifying the user that they will receive an email upon completion of differential expression analysis. This is because differential expression analysis using PUMA takes a long time, approximately 8 hours for our bovine dataset on a machine that boasts a 2GHz Intel CoreTM 2 Duo CPU with 4 gigabytes of RAM. This can be reduced by parallelising the differential expression process across multiple cores of a single CPU, or across multiple processing cores of multiple machines.

Differential expression analysis of dual and single dye data is similar to that of Affymetrix arrays, but because of the nature of the system naturally only requires one experimental design where the levels of the same factor are compared there is no need to specify the contrasts of interest.
contrasts, as there is only one contrast available. Processing options also differ significantly. Spots not on a dual array can sometimes be duplicated on an array. This is to give a more reliable measure of the expression level of a gene. The system allows for this by allowing the user to specify the number of duplicate spots per array and the number of space between duplicates. Spots that are not in the same row or column of duplicates are not duplicated and the duplicates may be partly spotted. This does not affect expression measurements. The level of correlation between these duplicate spots can then be factored into the linear model fit and subsequent differential expression analysis, hence giving more accurate results.

Gene annotation information can be dynamically downloaded by clicking the gene names in lists of differentially expressed genes (Fig. 4). This information can then be interpreted by the biologist. The .RData file or the spreadsheet listing differentially expressed genes can be downloaded if subsequent analysis is to be performed.

Figure 4: Screenshot showing differential expression analysis and clickable gene names.
The system makes a number of tools available to the user to manage data from previous experiments and analysis of experiments. The "Experiments" link in the menu on the left of the page will display a list of experiments that the user has previously uploaded and that are saved on the system. The option is available to remove any of these experiments from the system by checking them off and clicking the delete key. The page also allows the user to remove previously completed quality control analysis by clicking on the "Quality Control" link under the "Saved Analysis" menu. This page also allows the user to delete previous quality control analysis that was previously completed. The user may then select any previous experiments on the list for current use, which will allow them to either review previous analysis information or to perform new analyses. Clicking on any of the experiment names will bring the user to a page that shows various information about the dataset, such as the names of the files that were uploaded and their sizes.

Every quality control analysis is automatically saved by the system and can be re-viewed at a later date by the user. This is done by clicking the "Quality Control" link under the "Saved Analysis" menu. Once this is done, the user will be brought to a page with a list of all previous quality control analyses that have been completed for this experiment. This page also allows the user to delete these previous quality control analyses if desired by clicking on the delete key. The user may also access the condition under which the analysis took place, such as which arrays were selected, which preprocessing methods were used, and what phenotypic data was assigned.

If during quality analysis, the user had selected a PUMA PCA and screen plot, they will find those plots upon their completion. The user will receive an email informing them of completion, and the results can then be found under this menu. This is similar to the access of previous quality analysis. The user can delete previous analysis, access information on previous analyses, and view their results in full. The user also has the option of downloading the "RData" file, which is a saved version of the analysis results. The file may also be downloaded on the command line. The user may also choose to download the "RData" file and load it locally if they wish to pursue further analysis in the command line. Clicking the "More Info" link beside the link to download the "RData" file will provide the user with a detailed description of what objects are in the file and how they were created.
If the user has specified that analysis should be performed using PUMA, similarly to quality control analysis, the results will appear upon completion.
References


Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs and Terence P. Speed (2003), Summaries of Affymetrix GeneChip probe level data Nucleic Acids Research 31(4) e15


Publications arising from this project:
