

Criticality of Microarray Data Management

Introduction

The development of genomic and post-genomic technologies has created an explosion in the quantity, diversity and availability of both biological data and methods of analysis. Biologists are currently facing a dearth of efficient resources to convert this raw data into valuable knowledge. This paper presents a software platform designed to handle data and/or methods in the context of genome analysis.

Functional genomics can be defined as development and application of global (genome-wide or system-wide) experimental approaches to assess gene function by making use of the information and reagents provided by structural genomics.

Microarray technology is one of the most powerful functional genomics technologies that are used for measuring levels of gene expression. Microarray technology has revolutionized the study of gene expression, facilitating the monitoring of expression levels of thousands of genes simultaneously. This novel technique helps us understand gene regulation as well as gene-by-gene interactions more systematically.

Functional Genomics

Functional genomics is characterized by high throughput or large-scale experimental methodologies combined with statistical and computational analysis of the results. The fundamental strategy of a high throughput functional genomics approach is to expand the scope of biological investigation from studying a single gene or a protein to studying all genes or proteins of any given genome, at once, in a systematic fashion. Computational biology will perform a critical and expanding role in this area.

Structural genomics involves creating 'maps' of genomes, carrying out sequences, and determining the localization of genes and regulatory regions on chromosomes. Functional genomics studies the function (coded proteins), expression, and regulation of genes as well as the interactions between different genes. Functional genomics, thus, promises to rapidly narrow the gap between sequence and function, and to yield new insights into the behavior of biological systems.

Post-genome era data explosion

Genome sequencing projects generate complete records of the genetic make-up of organisms. These core data sets are complex, and pose challenges to those who seek to store, analyze and present the information. However, in addition to the sequence data, high throughput experiments like microarrays, SAGE etc., in combination with computing algorithms, generate distinctive new data sets like normalized data sets, clustered data sets, etc. The effective description and management of such data is of considerable importance to biologists and bioinformaticians in the post-genomic era.

The Challenge

DNA arrays have become the preferred method for large-scale measurement of gene expression. Expression profiling using DNA arrays is a step in the direction of functional characterization. There are several different implementations of the DNA array principle for expression measurement. Miniaturized devices, glass microarrays, and oligonucleotide chips are the most promising in terms of throughput, and should eventually allow simultaneous measurement of expression on the complete set of genes.

Data handling in these high throughput experiments is, in itself, a major issue: users can be quickly swamped by tens of thousands of measurements and may not be able to handle them, much less extract the information they contain. A well-organized and coherent software suite is necessary to move from raw data to corrected and normalized expression values. Also, imaginative statistical and representation tools must be implemented to allow the biologist to look at data in as many ways as possible, while integrating information available on the Internet for those genes. Data archiving is a related but distinct issue. Indeed, the same arrays are normally used by several groups, and by several individuals within a single laboratory/across various laboratories. It therefore makes sense to ensure that all the data, in a suitably standardized form, is available to each participant in some kind of laboratory notebook system. An amenable and effective database system is clearly essential, together with appropriate standardization to ensure that data from different laboratories can be compared.

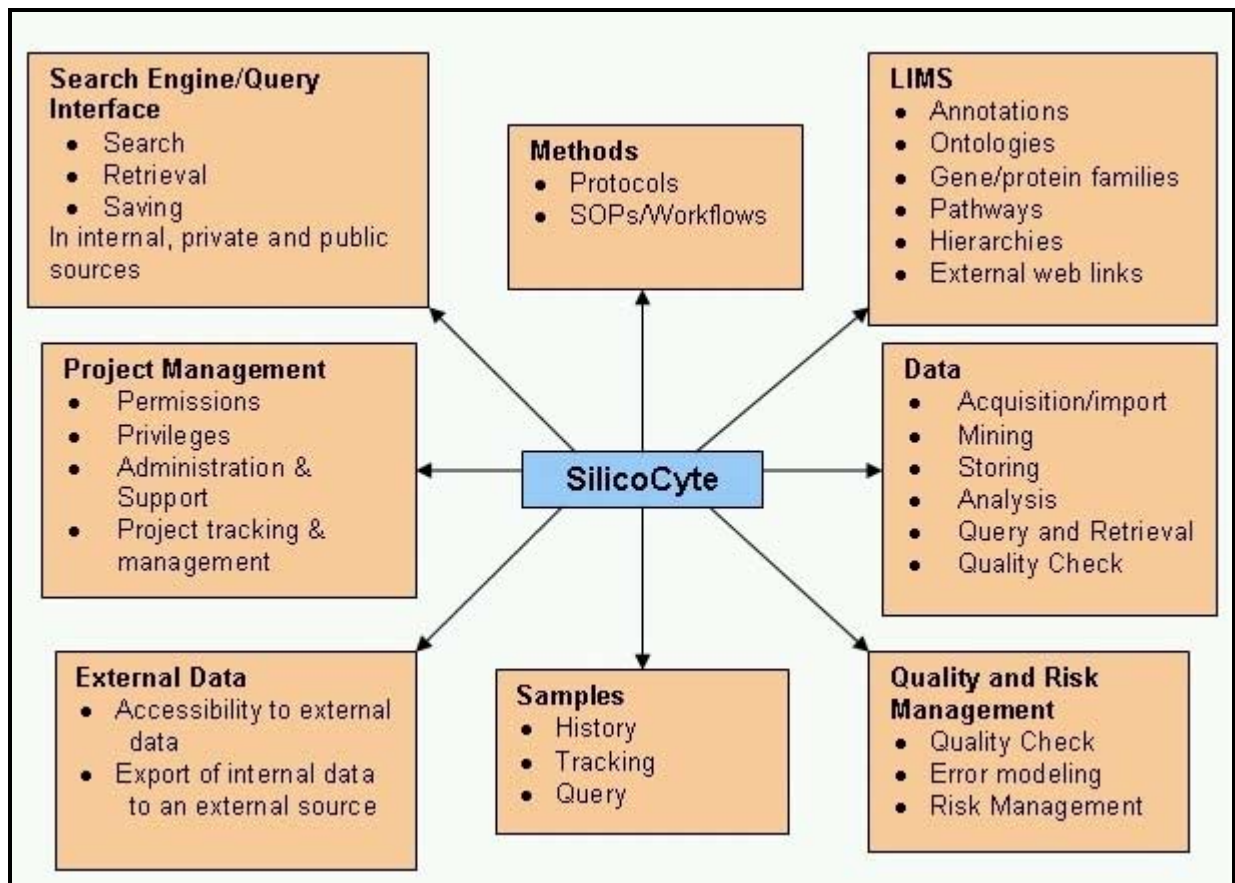
About SilicoCyte™

SilicoCyte™ provides an end-to-end solution for Microarray Data Analysis. It comprises of *Annotations*, *Image Analysis* and *Statistical Analysis* as core modules. SilicoCyte™'s data management system provides a viable solution to the challenges posed by incessant data generation in microarrays. It helps the user to acquire and process data by providing a smooth transition from assay performance to data analysis. Also, tracing and retrieving relevant data is made easy. A quality tab is kept on all the data that is generated in SilicoCyte™. Users can determine the quality of raw data and tailor subsequent analyses accordingly.

SilicoCyte™ provides the following features:

1. LIMS

- SilicoCyte™ allows the user to view, modify information about, and track genes, probes in the database.
- SilicoCyte™ includes Laboratory Information Management features like *Annotations, Hierarchical Information, Experiment Details* and *Project Management*.
- SilicoCyte™ provides links to various popular external links for connectivity between the information generated from the Experiments to public data repositories.
- SilicoCyte™ has various forms for navigating, deleting, and updating genes, gene groups, and hierarchies. You can also trace the gene to its source, import sources and source information.
- Organizing genes in multi-level tree based on various properties (for example, protein function, phylogeny, enzymatic function, or clone library) is possible. We currently provide three read-only main hierarchy branches - *Pathway Hierarchy, Function Hierarchy, Enzyme Hierarchy*. Custom hierarchies can also be created and maintained.
- External Web links on Hierarchies and Ontology are very useful for the user to correlate the data.



2. Data mining and analysis

Enormous amounts of gene expression data, that include comprehensive sample data and gene annotations, are generated. Such data must be managed efficiently. It needs to be integrated with data from customer's in-house research efforts, public domain genomic and medical database sources.

The data mining challenge is made easy for scientists with the aid of powerful analysis tools, such as in SilicoCyte™, for successfully generating and mining their own array-generated data to identify new therapeutic targets, develop improved diagnostic tests, and gain a better understanding of how biological systems work.

Our *Data Management and Analysis Software* is designed to manage, integrate, explore and query such disparate data sets as if they were part of a single database. SilicoCyte™ handles data from various sources and stores them in an organized manner and makes it available for analysis.

The single most important trend in data mining is the advent of statistics into microarray data analysis. An algorithm's strengths and weaknesses directly affect interpretations of results. Our statistical analysis tools are robust and user-friendly. Currently, SilicoCyte™ provides different data analysis algorithms for various statistical analyses like *Normalization, Dye-flip ANOVA, Replicate* and *Venn Analyses, Principal Component Analysis (PCA)* and *Clustering* techniques.

Microarray experiments can evidence lot of systematic variation, which affect the measured gene expression levels. Different sources, for e.g., differences in labeling efficiency between the two fluorescent dyes used in a two-channel microarray experiment, could be responsible for this variation. Normalization of raw data is the first step for any data analysis to eliminate systematic variation and should be performed both globally and locally with different parameters. There are eight normalization methods along with Lowess that are currently available in SilicoCyte. These are Normalization by mean, median, 10% trimmed mean, 75% percentile, scale between 0 and 1, subtract the mean, subtract the median, z score calculation. All these methods can be performed on all elements, pin/zone, rank and gene Groups. Lowess is a regression-based method of normalization to correct the dye bias across the entire range of signal intensities and pin or zone based.

Replicate analysis is another means to eliminate variation. It is required to perform statistical tests such as t-test, ANOVA, to compare one group to another. The Replicate Analysis feature in SilicoCyte facilitates the user to view the Mean, Median, SD, Variance, and Coefficient of Variation for the data pooled out from the replicates. This feature helps the user to filter data that produce erroneous results.

Principal Component Analysis (PCA) is a powerful data reduction tool in SilicoCyte. It can be used to characterize the most abundant themes or building blocks that reoccur in genes during Microarray experiments. PCA aims at reducing the dimensionality of data set without significant loss of information. This decomposition technique produces a set of expression patterns known as principal components; linear combinations of these patterns can be assembled to represent the behavior of all genes in a given data set.

ANOVA is used to obtain a confidence measure for differential expression of a gene between two conditions. SilicoCyte™ performs ANOVA data analysis on replicate data received from a dye-flip Experiment. To perform ANOVA there should be at least two groups, based on the sample taken. The samples could be treated vs. untreated, drug treated at hour one vs. at hour two etc.

Assumptions of ANOVA:

ANOVA comparison assumes that:

1. The data is (approximately) normally distributed.

2. The variances of the separate groups are almost equal.

If the data does not fulfill these conditions, the ANOVA comparison gives unreliable results.

Currently SilicoCyte performs one-way ANOVA that computes the difference between groups by comparing the mean values of the data in each group. The results are obtained by testing the above mentioned assumptions for null hypothesis, which states that there is no difference between the means of the groups. The result of one-way ANOVA is in the form of P-Value, which is the probability of the actual or more extreme outcomes under the null hypothesis.

SilicoCyte™ provides clustering of genes in three different ways viz., **Hierarchical Clustering, K-means Clustering and Self Organized Maps (SOMs)**. These clustering algorithms arrange genes according to similarity in pattern of their expression. The output is displayed graphically, conveying the clustering and underlying expression data simultaneously, in a form that is interesting to biologists. The clustering methods in SilicoCyte™ are combined with a graphical representation of the primary data. Here, each data point is shown in a color that quantitatively and qualitatively reflects the original experimental observations. The end product is a representation of complex gene expression data that, through statistical organization and graphical representation, allows biologists to assimilate and explore the data in a natural intuitive manner.

In **Hierarchical Clustering**, the data are partitioned in a hierarchical manner wherein a series of partitions take place. This may run from a single cluster containing all objects to n clusters each containing a single object. Hierarchical clustering is represented by a two-dimensional diagram known as dendrogram, which illustrates the fusions or divisions made at each successive stage of analysis. K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. It is well suited to generating global clusters. The K-Means method is numerical, unsupervised, non-deterministic and iterative. Self organizing maps (SOMs) is also a dimensionality reduction technique. It employs neural network concepts to achieve this.

Through SilicoCyte™'s **Visualization** engine, the users can look at the data they brought together, to explore relationships and trends by subsets and variables. The Visualization tool in SilicoCyte™ provides various graphical representation tools like *Scatter plots, 2-D and 3-D graphs* on both image and numerical data sets. Data Clustering and pattern matching have been integrated into SilicoCyte's searches. Gene expression patterns can be traced through multi-dimensional visual environments that are provided in SilicoCyte™.

3. Search engine/Query Interface

SilicoCyte™ provides a query interface where the user can query, retrieve and save data in a user-friendly format. This can be done on various datasets available in the database for efficient correlation across different experiments. This facility would help scientists to correlate the results across various laboratories for comparing the experiments to consolidate their findings.

4. Samples/Inventory

SilicoCyte™ can act as a repository for Sample and Experiment management and also tracking. It can store various protocols/SOPS/workflows that are generated in a particular laboratory and can efficiently retrieve them at the user's request.

5. External Data

SilicoCyte™ can import image analyzed or upstream data from a third party software for further data analysis. Currently, there are four types of image analyzed data that can be imported into SilicoCyte™, viz. Agilent, ScanArray Express (.csv), GenePix Pro (GPR) and Affymetrix data.

SilicoCyte™ supports *Analysis* and *Data Management* for dual channel c-DNA microarray for e.g. *Agilent, ScanArray Express, GPR data* and also single channel image analyzed data exported from *Affymetrix MAS Suite 5.0* and *GCOS 1.2*. This helps the user to perform data analysis on various data imported from third party software.

6. Quality and Risk Management

Microarrays as a means of high throughput screening of genes and potential targets results in enormous amount of data. This data could contain genuine outliers for both high and low signals due to dust or dropouts in the data. Also, variability is inherent in all array-based gene expression data. This natural variation is due to differences in how genes respond to the specific experimental conditions of the array. Automated detection of outliers and quantifying the variability in a production environment is a requirement in microarray analysis for accurate measurement of signals and in building high quality databases for further data mining.

Apart from standard validation procedures for the quality of RNA samples, labeling, and Hybridization for all GeneChip/microarray studies, it is important to address additional quality check (QC) issues during data analysis. SilicoCyte™ provides an extensive *QC Report* for *Image Analysis*, where graphical view of all array data along with its subarrays is represented. In addition, QC is performed on data by different normalization methods, p-value for every gene in case of Dye Flip ANOVA experiment, and replicate analysis. Statistical computations for image-analyzed data in the QC report include computation of entire array statistics and subarray statistics. This enables the user to take a decision on the

authenticity of the obtained results, and on whether to consider the array for further analysis. For example, if there is too much variation between the entire array statistics and subarray statistics the user can reject the experiment.

Expression data is visualized in the QC Report through graphical tools such as Scatter plot. These will help the user to identify the bias in the data and decide on the validity of the experiment. For e.g., a scatter plot graph with data widely scattered may signify more noise introduced in the experiment. In addition, SilicoCyte™ enforces some quality measures in mapping the relevant fields and ensuring certain important fields to be imported for successful analysis when data is imported from external sources e.g. Affymetrix.

7. Project Management

Hierarchies can be customized to organize probes, microarrays and projects in SilicoCyte™. *Project Hierarchies* are represented as a tree structure in SilicoCyte™. SilicoCyte™ offers various project privileges. Depending on the entitled privilege, you could view, edit, or have full control over the project information. Administrator has the maximum control over a project.

Advantages

SilicoCyte™ deals with very large Data Sets and enables exploratory Data Analysis. It is a comprehensive tool for *Microarray Data Analysis*, right from *Annotations* to *Statistical Analysis*. It addresses the entirety of microarray experimentation, from the upstream stages of microarray design to the downstream stages of post-processing. SilicoCyte™ highlights interesting patterns, thus helping biologists to arrive at better conclusions and consolidate their findings. An end-to-end solution tool like SilicoCyte™ brings in better understanding of various biological and biochemical pathways and processes to aid a researcher in identifying right candidate genes, proteins and drug targets, reformulate their hypotheses and to know better which questions to ask.

SilicoCyte™ provides a comprehensive solution to managing and mining microarray data. It brings together a variety of mathematics and database tools to help ensure that the researchers can access the right analytical methods.