XMAS: An Experiential Approach for Visualization, Analysis, and Exploration of Time Series Microarray Data*

Ben Dalziel[1], Hui Yang[1,**], Rahul Singh[1,**], Matthew Gormley[2], and Susan Fisher[2,3,4,5,6]

[1] Department of Computer Science, San Francisco State University
huiyang@cs.sfsu.edu,rsingh@cs.sfsu.edu
[2] Departments of Cell and Tissue Biology
[3] Anatomy
[4] Obstetrics, Gynecology and Reproductive Sciences
[5] Pharmaceutical Chemistry
[6] Institute for Regeneration Medicine and the Human embryonic Stem Cell Program, University of California San Francisco. San Francisco, CA

Abstract. Time series microarray analysis provides an invaluable insight into the genetic progression of biological processes, such as pregnancy and disease. Many algorithms and systems exist to meet the challenge of extracting knowledge from the resultant data sets, but traditional methods limit user interaction, and depend heavily on statistical, black box techniques. In this paper we present a new design philosophy based on increased human computer synergy to overcome these limitations, and facilitate an improved analysis experience. We present an implementation of this philosophy, XMAS (eXperiential Microarray Analysis System) which supports a new kind of "sit forward" analysis through visual interaction and interoperable operators. Domain knowledge, (such as pathway information) is integrated directly into the system to aid users in their analysis. In contrast to the "sit back", algorithmic approach of traditional systems, XMAS emphasizes interaction and the power, and knowledge transfer potential of facilitating an analysis in which the user directly experiences the data. Evaluation demonstrates the significance and necessity of such a philosophy and approach, proving the efficacy of XMAS not only as tool for validation and sense making, but also as an unparalleled source of serendipitous results. Finally, one can download XMAS at http://cose-stor.sfsu.edu/~huiyang/xmas_website/xmas.html

1 Introduction

Microarray-based experimentation is a technique, which measures the expression levels for hundreds and thousands of genes within a tissue or cell simultaneously. It

[*] This research was partially funded by a grant from the National Science Foundation IIS-0644418 (CAREER) – Singh; SFSU CCLS Mini Grants – Yang.

[**] Corresponding authors.

M. Elloumi et al. (Eds.): BIRD 2008, CCIS 13, pp. 16–31, 2008.

therefore provides a data rich environment to obtain a systemic understanding of various biochemical processes and their interactions. Data from microarray experiments have been used, among others, to infer probable functions of known or newly discovered genes based on similarities in expression patterns with genes of known functionality, reveal new expression patterns of genes across gene families, and even uncover entirely new categories of genes [1], [2]. In more applied settings, microarray data has provided biologist with ways of identifying genes that are implicated in various diseases through the comparison of expression patterns in diseased and healthy tissues.

The area of microarray data analysis remains particularly active, leading to the development of numerous algorithms and software tools. The algorithmic underpinnings of these methods span a variety of pattern analysis, machine learning, and data mining methodologies including Bayesian belief networks (BBN), clustering, support vector machines (SVM), neural networks and Hidden Markov models. A survey of many of these techniques can be found in [1], [3], [4] and [5]. From a user perspective, a number of vendors have developed software systems for microarray data analysis such as Ingenuity [6], Onto-Express [7] and GenMAPP [8]. Furthermore, plug-ins have been developed for existing software systems such as the BioConductor [9] package for R [10], along with SAM [11] and PAM [12] for Excel.

Despite this un-arguable richness of analysis tools, it is acknowledged however, that analysis of microarray data is currently at a bottleneck [13]. Some of the most fundamental reasons behind this include:

- *Emphasis on the algorithmics to the exclusion of the user*: Holistically taken, most microarray analysis implementations are algorithm-oriented and do not provide sufficient support for exploration and/or hypotheses formulation. From an end user perspective, they function as a "black box" giving users very limited control over the analysis process outside what the underlying algorithmic mechanism is intended for. Among others, this limits the ability of users to integrate their domain expertise into the analysis process or explore alternatives which the algorithm design had not foreseen.
- *Interpretability*: Methods involving complex algorithms (such as BBN, SVM, and dimensionality reduction) may produce results that are difficult to interpret or understand. This can create a disconnect between the algorithmic process and the biochemical interpretability of the information.
- *Biased statistical analysis*: An important challenge outside the aforementioned user-centric issues lies in the fact that many existing techniques (e.g., SAM and PAM) employ statistical approaches to analyze microarray data. This can lead to bias, since in the majority of microarray studies the data is under-constrained (there are far fewer samples than genes or probes of interest). A representative example is the dataset used in this paper. It studies the placenta over the duration of pregnancy and is composed of just 36 samples containing expression levels for over 40,000 probes. As a result, it is difficult to construct reliable statistical samples or assume a reasonable data distribution model to carry out further analysis.

Given the aforementioned context, we propose re-thinking the design philosophy for developing microarray data analysis systems. Our central observation notes the fact that computers are inherently strong at large scale processing, data storage and data integration. However they lack the human skills of contextual reasoning, pattern detection, hypotheses formulation, exploratory behaviors, and sense making. Thus the primary design goal we seek to establish is the ability to exploit human-machine synergy by taking advantage of the aforementioned complementarities.

In the area of human-computer interactions, such an emphasis on exploration and hypothesis formulation in data rich environments has been the focus of study in [14] and [15], where the term "experiential environment" was used to denote systems and interfaces that take advantage of the human-machine synergy and allow users to use their senses and directly interact with the data.

In this paper, we describe the anatomy of a microarray data analysis system called XMAS (eXperiential Microarray Analysis System) that is developed by using and extending the ideas of experiential computing. The proposed system is (1) direct in that it does not use complex metaphors and commands; (2) supports unified query and presentation spaces; (3) maintains user context; (4) provides external contextual information through assimilating a variety of supplementary data such as pathway data from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [16], and (5) supports algorithmic and user-directed analysis, exploration and hypotheses formulation. Our ultimate goal is to promote perceptual analysis by integrating the user directly in an interactive and reflexive visualization environment with powerful algorithmic capabilities. XMAS is not limited to the analysis of time series microarray data, and can be more widely applied to any time-series datasets. XMAS supports the following visualization and analyses:

- *Trajectory based gene clustering*: In time series microarray data, a trajectory is composed of a sequence of expression measurements collected at different time points for a certain probe or gene. It is essentially a time series of gene expression data w.r.t. a single probe or gene. This function clusters different genes according to the relative geometric similarity of their expression trajectories.
- *Data filtering*: This can be based on gene identifiers, pathways, and integrated or user defined annotations. These filters facilitate the specification of genes of interest, enabling the user to narrow down hypotheses. This functionality extends to support any integrated secondary data.
- *Interestingness evaluators*: XMAS implements a set of measurements such as Pearson's correlation and p-value to quantify the interestingness of the results, to aid the user during visual inspection and more generally the entire analysis process.
- *Visualizations*: Two primary visualizations provide interactive representations of data at different resolutions including (1) a discretized trajectory view; and (2) a precise gene expression view.
- *Interactions*: Users can directly manipulate, interact and explore the data using highly intuitive point-and-click interactions.

There exist systems which support some of the features described above. For example the commercial system OmniViz [17] offers various reflective and interactive visualizations in addition to the more traditional statistical measures and algorithmic capabilities. Systems which share this closer resemblance to XMAS lack the core experiential design philosophy, which in turn has a significant influence over the completed system in the following areas: interaction, visualization, data integration, and interoperability. This will become apparent through the remainder of the paper.

Through use of XMAS, users are expected to achieve three main goals: (1) to gain a deeper understanding of a time series microarray dataset; (2) to verify or compare phenomena reported in literature on comparable datasets; (3) to generate hypotheses through examining results from different analyses. The main contributions of this work include: (1) increased user involvement, comprehension and understanding through development of a new design philosophy for microarray data analysis; (2) improved biological results from analysis; and (3) a concrete web-based extensible implementation of this design philosophy. This paper goes on to describe XMAS; its fundamental components and associated combinatorial power in Section 2. In Section 3 experimental results and user evaluation are presented to demonstrate the efficacy of this approach.

2 System Description

XMAS is an experiential system for time series microarray data (TSMAD) analysis through realizing a collection of interactive visual data operators and assimilating different types of knowledge such as pathway information. As shown in Fig. 1, XMAS consists of the following main modules: (1) data preprocessing; (2) a collection of interoperable data operators, including a parameterized discretization operator, basic data integration operators, and trajectory-oriented data operators; (3) interestingness evaluators; and (4) visualization and Human Computer Interaction (HCI). Next, we first discuss the datasets utilized by XMAS, and then describe in detail its main modules.

2.1 Data Sets

XMAS focuses on the analysis of time series microarray data. Such data has been used to study the developmental nature of an organ (e.g., a cancerous tissue) by conducting Microarray experiments on samples drawn from this organ over time. The

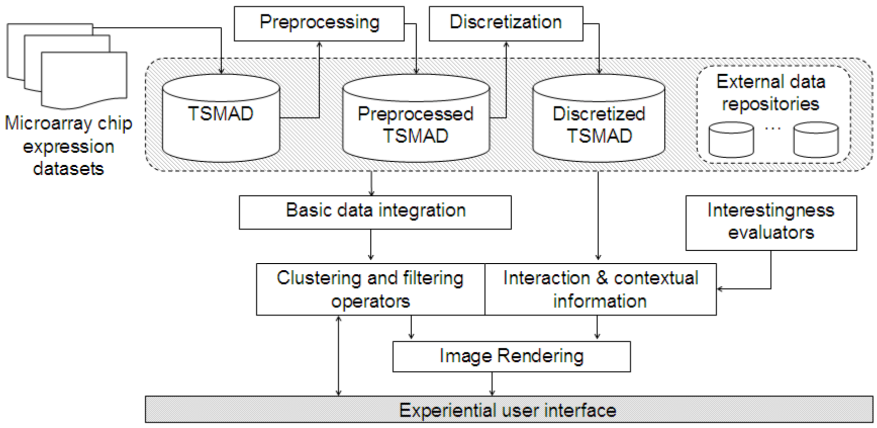

Preprocessing
Discretization
Microarray chip expression datasets
TSMAD
Preprocessed TSMAD
Discretized TSMAD
External data repositories
...
Basic data integration
Interestingness evaluators
Clustering and filtering operators
Interaction & contextual information
Image Rendering
Experiential user interface

Fig. 1. System overview of XMAS

genes of interest are generally specific to a study, which in turn determines the set of probes on a microarray chip that one is interested in looking into.

Let $\boldsymbol{D}$ be a TSMAD, $P=\{p_1, p_2, .., p_M\}$ be the set of M probes of interest, $T=<t_1, t_2, .., t_Q>$ be the ordered Q time points when Microarray analyses are conducted, and $S_i=\{ s_1, s_2, .., s_{Ni} \}$ be the set of N_i samples at *time* $t_i \in T$. Note that $\mathrm{S_i}$ and $\mathrm{S_j}$ (i≠j) might be two different sets due to restrictions on acquisition of live tissues. Then $\boldsymbol{D}$ can be considered as a dataset of M time series, each of which corresponds to one probe and is referred to as a *complex probe trajectory*. For each probe $p_k \in P$, its trajectory has Q time points. Each time point is associated with a vector of N_i expression values, corresponding to the N_i samples at this point. To further enhance users' explorative power, and analysis experience, XMAS integrates a variety of existing domain knowledge such as a mapping database between the probe set P and the set of genes, and pathway data from KEGG. XMAS adopts MySQL, an open source RDMBS, to manage such data.

2.2 Data Preprocessing

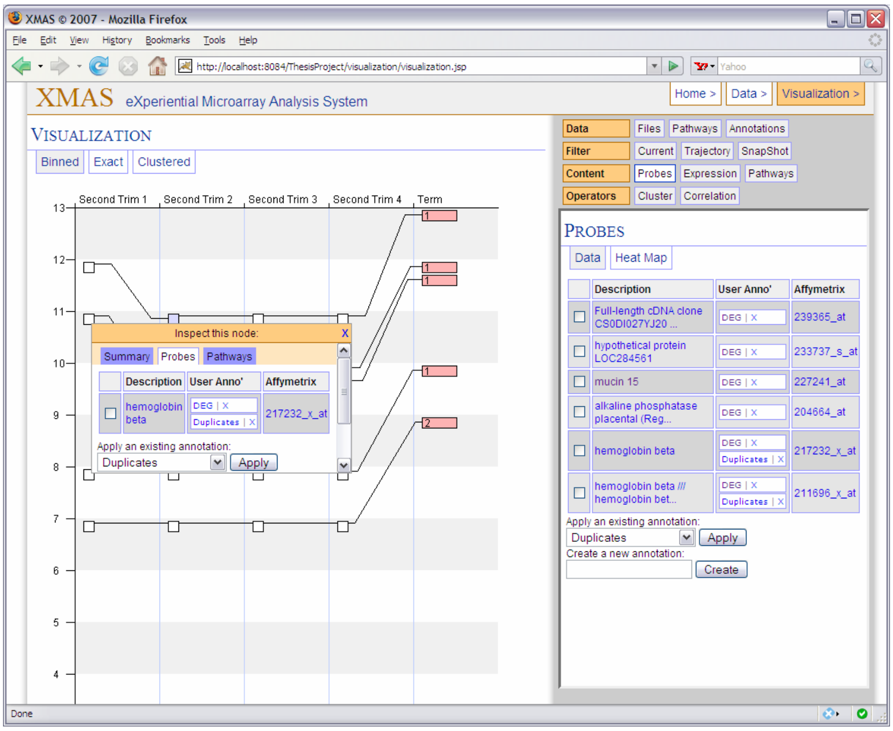
Given a TSMAD $\boldsymbol{D}$, this module first performs a base-2 logarithmic transformation over each expression value in $\boldsymbol{D}$. It then applies a simple data reduction technique to reduce each complex probe trajectory to a simple time series. Specifically, for a given complex trajectory, it replaces the vector of expression values at each time point by the median of this vector. One main reason the median is chosen is that it is more noise-tolerant. For the remainder of this paper, we refer to such simple time series as *simple probe trajectories* or *probe trajectories*. This process simplifies analysis at a global level, where the median expression is a reasonable representation of the constituent samples. Complete expression levels are preserved within XMAS and are accessible to aid in more concentrated analysis.

2.3 Interoperable Data Operators, Visualization and HCI

Interoperable data operators, intuitive visualization, and user-friendly HCI support form the core of XMAS. XMAS consists of data operators that can both function individually and collaborate with others when combined at users' command. Unlike most existing software systems for Microarray data analysis, XMAS injects visualization and HCI into data analysis. Therefore, users can not only visually observe the results at any moment, but also be able to interactively respond to XMAS to design their own explorative paths towards concept validation or hypothesis generation. It is due to this tight coupling of data operators, visualization and HCI, we will describe each data operator by also including the other two aspects.

Parameterized data discretization: One main interest in studying TSMADs is to characterize the temporal movement of genes in terms of expression level. Given that the collection of genes under study can be large, for instance, in the order of tens of thousands, examining a dataset on a trajectory-by-trajectory basis is time consuming and difficult. In addition, one also needs to reduce the impact from noise in the data. To address such issues, XMAS first applies equi-width discretization to each probe contained within the preprocessed TSMAD, where the width w (applied globally) is a user-specified parameter. The result of this intuitive probe association operator is a collection of *discretized probe trajectories*, where each expression level is represented by an integer, corresponding to its discretized value. The issue of information loss inherent to such discretization is countered through the preservation of the precise expression values which can be exposed through visualization or inspection.

Fig. 2 shows part of a screenshot of such discretized trajectories. In this figure, each discretized value (or bin) occupies one row space. Small squares or nodes in each bin can be clicked to reveal all the probes whose expression levels fall into this bin at a give time point. Moreover, all the nodes are arranged from left to right in columns, with the i^{th} column corresponding to the i^{th} time point. A node is colored in red if its expression level is higher than the previous node on a trajectory and blue if it is lower. The probes in the first node in a row share discretized expression value at the first time point. The probes in each of the rightmost nodes share identical discretized trajectories. And the probes in each of the middle nodes share a partial trajectory prior


XMAS © 2007 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8084/ThesisProject/visualization/visualization.jsp
Yahoo
XMAS eXperiential Microarray Analysis System
Home > Data > Visualization >
VISUALIZATION
Binned Exact Clustered
Second Trim 1
Second Trim 2
Second Trim 3
Second Trim 4
Term
13
12
11
10
9
8
7
6
5
4
1
1
1
1
2
Inspect this node:
X
Summary Probes Pathways
Data Files Pathways Annotations
Filter Current Trajectory SnapShot
Content Probes Expression Pathways
Operators Cluster Correlation
PROBES
Data Heat Map
Done

	Description	User Anno'	Affymetrix
□	hemoglobin beta	DEG | X Duplicates | X	217232_x_at

Apply an existing annotation:
Duplicates Apply

	Description	User Anno'	Affymetrix
□	Full-length cDNA clone CS0DI027YJ20 ...	DEG | X	239365_at
□	hypothetical protein LOC284561	DEG | X	233737_s_at
□	mucin 15	DEG | X	227241_at
□	alkaline phosphatase placental (Reg...	DEG | X	204664_at
□	hemoglobin beta	DEG | X Duplicates | X	217232_x_at
□	hemoglobin beta /// hemoglobin bet...	DEG | X Duplicates | X	211696_x_at

Apply an existing annotation:
Duplicates Apply
Create a new annotation:
Create

Fig. 2. The XMAS analysis environment is divided into three primary regions: (1) the visualization space displays discretized or precise trajectory views. Visualizations in this space can be manipulated in a similar way to various interactive web based mapping applications. This accommodates larger visualizations than would be practical in a static environment. Each node in the primary visualization is interactive, allowing the user to inspect content through in-place context windows (2). A complementary view, the visualization sidebar (3), provides similar data for the entire visualization. Operator specification tools in addition to operator summaries and correlation data are also accessible from this space.

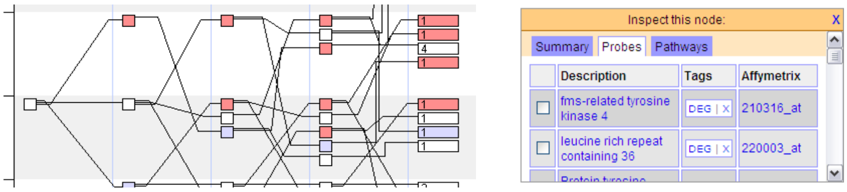
to that time point. All such nodes are expandable. Note that the system calls several operators described later to construct those nodes.

Basic data integration operators: The operators contained within this category realize integration of different datasets. They can be categorized as follows:

- *Gene-probe integrators:* These operators relate probes to genes or vice versa, for instance, identifying the list of probes associated with a given gene.
- *Probe-gene-pathway integrators:* This set of operators enriches a gene or probe with pathway information. For instance, one such operator determines whether a given gene participates in a pathway; whereas another operator lists all the genes or probes that are involved in a pathway.
- *Trajectory-trajectory integrators*: These operators relate the three forms of probe trajectories utilized by XMAS: complex, simple and discretized probe trajectories.

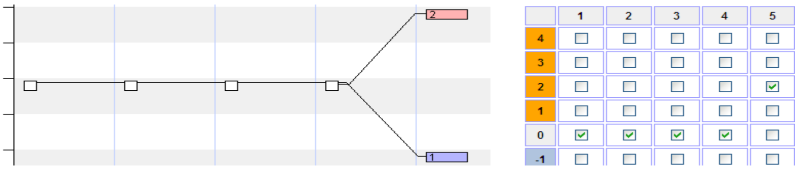
Trajectory-oriented data operators: This set of operators support users to explore the data by examining and uncovering the similarity among probe trajectories.

- *K-means clustering*: This operator puts probes of similar, non discretized trajectories into the same group. The user can choose to cluster based on either Euclidian or Pearson's Correlation distance metrics and can specify the value of K.
- *Expression level preserving trajectory-based clustering:* This operator identifies the genes whose discretized probe trajectories are identical and associates them in a single cluster. Two trajectories are identical if they have the same expression level at each time point. Fig. 2 shows examples of such clusters, each corresponding to one trajectory. One can inspect the probes and related contextual information in a cluster by clicking the corresponding node.
- *Trajectory shape based clustering:* This operator finds similar shaped trajectories across possibly different expression values. Probes of the same trajectory shape are essentially co-expressed at each time point. Therefore, each of such clusters identifies one co-expression pattern. We implement this operator in two steps. It first vertically translates all the discretized probe trajectories in a way such that the first node of each trajectory corresponds to the same expression level 0. For instance, for a given trajectory <2, 3, 1, 3, 4>, its translated trajectory is <0, 1, -1, 2>. The second step finds such clusters by calling the previous clustering operator. Fig. 3 shows part of a screenshot of such clusters. One can view the content of each cluster by expanding each of the rightmost nodes.



	Description	Tags	Affymetrix
□	fms-related tyrosine kinase 4	DEG | X	210316_at
□	leucine rich repeat containing 36	DEG | X	220003_at
	Protein tyrosine		

Fig. 3. Trajectory shape based clustering translates trajectories to a common root. Each node is interactive, revealing contextual data about the content of the node as a mobile, in-place window in the visualization.



	1	2	3	4	5
4	□	□	□	□	□
3	□	□	□	□	□
2	□	□	□	□	☑
1	□	□	□	□	□
0	☑	☑	☑	☑	□
-1	□	□	□	□	□

Fig. 4. Shape based trajectory specification, reveals 2 clusters of inverse trajectory shape

- *Discovery of inversely expressed probes/genes*: This operator identifies probes whose discretized trajectories are the inverse of each other. Fig. 4 shows the interactive query space and corresponding visualization showing five probes expressing with perfect discretized inverse correlation to twelve others.
- *Filtering operators:* Such operators utilize one or more basic data integration operators described earlier to identify trajectories that satisfy certain specified criteria. All such operators are integrated into one interactive user interface as shown in Fig. 2. XMAS currently supports the following filtering operators:
  - *Filtering by probes or genes*: This identifies probe trajectories associated with one or more specified genes.
  - *Filtering by pathway*: This identifies probes involved in a specified pathway
  - *Filtering by gene expression movement*: This identifies probes that are partially or entirely co-expressed. Fig. 4 illustrates the interface where users can specify a specific co-expression pattern of interest. This filtering operator can be applied to strictly trajectory-based clusters, or trajectory shape based clusters, as illustrated in Figs. 2 and 4 respectively. In Fig. 2, a user is interested in identifying all the probes or genes with a relative movement of 2 between the last two discretized expression levels. Fig. 4 illustrates the ability to include all the inversely expressed genes, this time for shape based clusters (i.e. with the same root). A similar operator is also included where one can identify the probes that have a similar expression level at one or more time points by specifying the range of expression levels at such time points.
  - *Exclude a probe from the resulted probe set*: This operator removes a probe from analysis. In Fig. 3, one can remove a probe by clicking the ‘x’ symbol.

Note that all the above data operators are interoperable with each other. This is essential, as XMAS does not prescribe data discovery paths for users. Instead, it empowers users to construct their own discovery paths by combining different operators in different order. XMAS achieves this by accommodating an integrated user interface shown in Fig. 2.

2.4 Interestingness Evaluators

Although visualization is powerful and intuitive for users to gain insight into a dataset, its effectiveness can be greatly reduced in a variety of situations. For instance, the amount of the data being visualized is too large to fit into a computer screen. In some cases, data might exhibit an inherently complex structure such that it is difficult for human beings to make sense of the visualized data. To overcome this limitation, XMAS includes a collection of evaluators to quantify the results.

- *Volatility of a trajectory*: Let $TR=<e_1, e_2, \ldots, e_Q>$ represent a discretized trajectory, where e_i is the expression value at the i^{th} time point. The volatility of this trajectory is defined as $\Sigma_{i=1,Q-1}(/ e_i - e_{i-1} /)$. One can use this measure to identify probes with extremely low or high volatility, where the former might not be of much interest and the latter might be a result of noise in the dataset.
- *Precision and recall*: These two measurements are used to quantify the strength of association between a pathway and the set of probes produced by a data operator. Let P be the pathway of interest and x be the number of participating probes of P. Let y be the number of probes returned by a certain operator, among which z probes are associated with P. Then Precision=z/y and Recall=z/x.
- *Pearson's correlation coefficient*: Let $X=<x_1, x_2, \ldots, x_Q>$ and $Y=<y_1, y_2, \ldots, y_Q>$ be two probe trajectories. One can use this evaluator to measure the direction and strength of the linear relationship between *X* and *Y*.
- *Identification of differentially expressed genes (DEGs)*: DEGs are selected by determining the moderated t statistic-adjusted P values (<0.05 using Bonferroni correction [18]). Fig. 2 highlights the DEGs within the current analysis, as leaf annotations in the primary visualization (1), and as "tags" in the list view (3).

P-value: We adopt P-values to measure the statistical and biological significance of observing a set of probes being associated with each other by a clustering operator described earlier. Given a background distribution, the lower the p-value, the more unlikely that observing a set of probes associated with each other is by chance. We next use the pathway annotation as an example to explain how P-values are computed. Let *N* be the number of probes under study, *D* be the number of probes in a given pathway, *n* out of these *N* probes are associated with each other by a data operator, and finally, *k* out of these *n* probes are also in the said pathway. The P-value of this association of n probes is then defined as: $P - value = \binom{D}{k}\binom{N-D}{n-k} \quad \binom{N}{n}$

2.5 Interoperability, Interactivity and Extensibility of XMAS

Interoperability among data operators: Unlike most existing software tools for TSMADs, XMAS does not prescribe analytical tasks for users. Instead, it empowers users to construct their own data discovery paths tailored for their special needs by combining different operators in different orders. XMAS achieves this by realizing interoperable data operators and an integrative user interface shown in Fig. 2. Aided by visualization, users can use this interface to select a sequence of data operators that are most likely to maximize their understanding of a problem at hand. A use case is described in detail in section 3 to illustrate this feature and its advantages.

Interactivity: Interactions with operators in XMAS are direct, i.e., no complex metaphors are involved. In addition, XMAS maintains contextual information on both users' behavior and data produced from such behavior. This ensures that there is no unnecessary context switching, thereby reducing the cognitive load from users.

Extensibility: Due to its modular architecture design (Fig. 2), XMAS can be readily extended in one or more of the following aspects: (1) integrate additional supplementary datasets such as gene ontology (GO)[19] functional categories and implement new data integration operators to enrich users' analytical experience; (2) integrate new data operators; and (3) realize additional interestingness evaluators.

3 Experimental Evaluation

XMAS' analytical power lies in the union of three areas: (1) visualization; (2) interactivity; and (3) interoperability. As discussed earlier, existing algorithms and software systems lack some or all of these desirable components. Considering the general trends in the state of the art: user interaction is limited to data entry, parameter specification and analysis via a simple (text based) command driven interface. Workflow is linear and disjoint, (often spread over numerous systems), and data presentation is generally textual (with notable exceptions such as pathway visualization in GenMAPP).

In this section we present evaluation of XMAS which demonstrates the importance and necessity of having the three areas coexist. First, we describe how XMAS can be used as an interactive visual tool to foster a greater breadth and depth of understanding within microarray data. Second, a common information goal serves as the entry point to a highly-non-traditional workflow drawing on many interoperable components of XMAS. Finally, comparative quality information is presented to support the generated hypotheses. Throughout, the inherent facilitation of hypothesis generation and serendipitous discoveries are highlighted. All evaluations were performed on the data set described below.

3.1 Data Description

To demonstrate the efficacy of XMAS, we used it to analyze a publicly available TSMAD [GEO Accession No: GSE5999] which captures expression data of human placentas during pregnancy. Using the description of a TSMAD provided in section 2.1, five time points (Q=5), comprising N_1=6, N_2=9, N_3=6, N_4=6, and N_5=9 samples capture genome wide (45,000 probes representing 39,000 gene transcripts) expression profiles of non-contiguous placentas between 14 and 40 weeks of pregnancy. The 5 distinct gestational time intervals (Q) range between 14-16, 18-19, 21, 23-24, and 37-40 weeks. The experiments which compose Q=1 through Q=4 capture the stage of pregnancy known as midgestation, and the samples from Q=5 are contained within the third trimester, also known as Term. For complete experimental protocol, description and analysis workflow, readers are referred to [20]. The findings on this dataset, reported in [20] will be cross-referenced where necessary. The dataset was first preprocessed as described in Section 2.2. It was then discretized as explained in Section 2.3. Throughout the following evaluation, a bin size of 1 (i.e., w=1) was used.

3.2 XMAS as a Visual Interactive Tool to Aid in Data Comprehension

Developing a detailed understanding of a TSMAD is an important step towards generating focused analysis and hypotheses. Traditionally, the development of a broad and formal understanding is based almost exclusively on the dissection of output from utilizing a variety of analysis systems and algorithms. In contrast, XMAS provides an integrated environment to facilitate this process. We next describe two scenarios (among many), where XMAS is being used to help expert users gain both a global and localized view of the data and many times serendipitous discoveries.

Expression pattern knowledge discovery**:** Visualization of discretized trajectories and shaped based trajectory clustering (i.e. unique trajectories) provide a global view of the entire dataset (Fig. 2(1)). As the user began to specify the operators (Section 2.3), the reflective query space updated to indicate the quantity of probes, DEGs and unique trajectories that would match the defined operator (Fig. 2). This reciprocal interaction aided the user to gain insight into the distribution of probes, DEGs, and the variability of probe expression during the specification refinement process. For instance, with 2 mouse clicks—one for the discretization operator and the other for the shape-based trajectory clustering--XMAS reveals that there are 76 distinct expression patterns and 504 DEGs in the dataset. Using the filter as shown in Fig. 4, more detailed information of such patterns were identified within a few mouse clicks: 6 patterns showing a significant expression increase (≥ 4-fold) at Term, 11 showing an expression decrease (≥4-fold) at Term, and only 1 showing a 16-fold increase. One more click revealed that only one probe involved in the last case. Such information provides the user with an insight into both the global *and* localized behavior of their data. This is in sharp contrast to traditional analyses, where such information is gleaned through utilizing a number of tools. Additionally, due to effective integration of user knowledge, our evaluation has shown that XMAS can often uncover previously unknown, yet interesting patterns in the data, thereby leading to serendipitous discoveries.

Pathway involvement analysis**:** The identification of known biological processes (or pathways) involved in a TSMAD is one main goal in microarray analysis. Following the identification of such pathways, domain users often find it necessary to further support such identification by investigating the relative involvement of each pathway in the context of the entire data set (i.e. not exclusive to DEGs, which are traditionally the sole focus of pathway analysis such as GenMAPP). This is generally a labor-intensive and manual process, which can take up to several hours and may become impractical for large pathways. We next use the *Apoptosis* pathway as an example to demonstrate how XMAS can significantly improve in this respect.

As illustrated in Fig. 5, we first used the pathway membership filter to identify the 631 probes involved in the *Apoptosis* pathway, among which 8 were annotated as DEGs. We then inspected the annotations accompanying each discretized trajectory in the visualization, to ascertain the quantity of probes sharing DEG expression profiles (at the discretized level). This, the user determined, was a good way of assessing the relative involvement of the entire pathway. Individual probes were subsequently re-included into analysis, enabling visual assessment on a probe-by-probe basis.

This simple concatenation of operators led to a focused analysis of pathway involvement, reducing what was previously a multi hour process to a few interactions (mouse clicks). Too often, traditional analysis concentrates exclusively on DEG lists, and here, simple trajectory association enabled the user to surround DEGs with

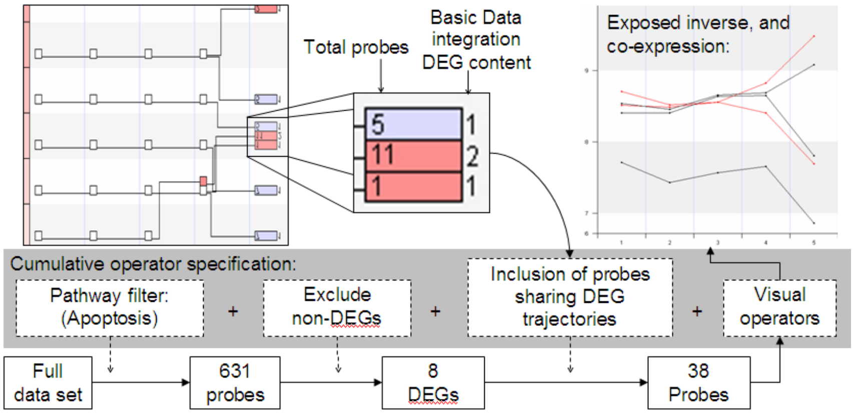

Total probes
Basic Data integration DEG content
5
11
1
1
2
1
Exposed inverse, and co-expression:
Cumulative operator specification:
Pathway filter: (Apoptosis)
+
Exclude non-DEGs
+
Inclusion of probes sharing DEG trajectories
+
Visual operators
Full data set
631 probes
8 DEGs
38 Probes

Fig. 5. User led analysis quickly identified DEG involvement in a given pathway. Probe context information presented within the visualization enables the user to pull in similar, yet non-DEG probes to focus an analysis on the relative involvement of the pathway as a whole.

contextually similar probes. Analysis of these probes facilitated a more confident declaration of significance, and led to the specification of a subset of probes which could form the basis of subsequent analysis.

3.3 Negative Expression Shift Approaching the End of Pregnancy

In this and the following sections, we described a complete workflow to illustrate the power of discovering serendipitous knowledge as a direct consequence of the integration of visualization, interactivity, and interoperability among data operators. Such integration enables a highly focused, yet simple analysis, which leads to the exposure of pathway involvement, hypothesized crosstalk, and co-expression patterns. These types of knowledge could not be reasonably developed by traditional means. The user workflow is described below and illustrated in Figs. 6 and 7.

Towards the end of pregnancy, the placenta begins to shut down in preparation for delivery. This process materializes at the genetic level as placental cells switch off, and is observed as a shift in expression between the second trimester intervals and term (time period 5) 0. The entry point to this analysis was to identify such probes.

Traditionally, such analysis involves the reduction of the data set into two representative samples, between which the expression characteristic can be evaluated. However, considerable details can be lost in this process. The analysis from 0, for example, assumed constant expression during midgestation, reducing 27 samples to just one. This is not the case, as one can observe directly within XMAS (Fig. 3). Furthermore, the lack of interaction in traditional analyses heavily restricts the users' ability to obtain a greater sense of completeness.

As shown if Fig. 6, we first performed a trajectory shape-based clustering to identify 39 probes that show a 4-fold or more increase at Term, of which 19 are DEGs. The visualization based contextual information further verified that the clustering

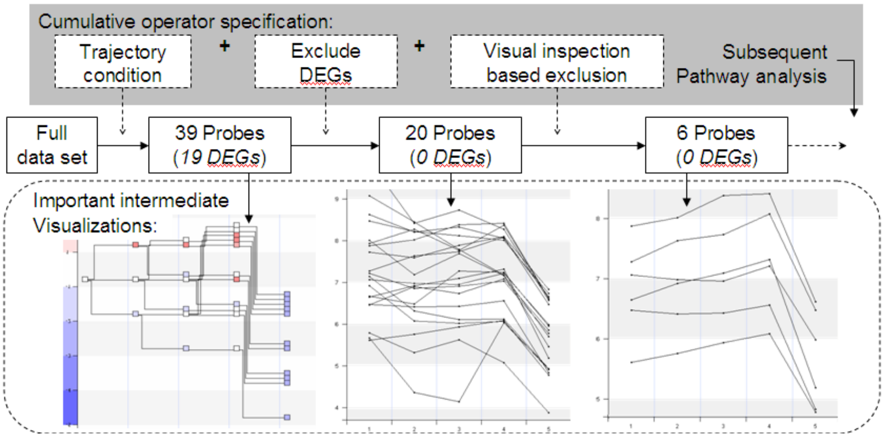

Cumulative operator specification:
Trajectory condition
+
Exclude DEGs
+
Visual inspection based exclusion
Subsequent Pathway analysis
Full data set
39 Probes (19 DEGs)
20 Probes (0 DEGs)
6 Probes (0 DEGs)
Important intermediate Visualizations:

Fig. 6. Workflow, illustrating the specification of operators used to focus analysis. Correlation scores demonstrate the power of exploratory analysis to expose common patterns of biological interest based on simple interactions. Key visualizations provide an insight into the environment in which the user is empowered to apply domain knowledge.

captured the target characteristic well. DEGs were subsequently excluded from analysis to concentrate on the remaining 20 candidates, as they share similar expression patterns with DEGs yet not categorized as DEGs. Through visual inspection of precise probe trajectories the user was able to exclude probes judged to be of lesser interest in the context of the current analysis. Interactively, we focused on the emergence of a specific trajectory shape, shared by 6 probes. Correlation analysis verified and strengthened this association. Through this process (Fig. 6), XMAS enables direct application of domain knowledge and intuition from the domain user. This is unmatched by other systems.

Main Observations: The quantity of discretized trajectories represented by the 39 probes (Fig. 6) indicates the details lost in traditional methods. XMAS facilitated a less strict, more intuitive specification of characteristics, which accommodated a greater sense of completeness than traditional analysis is capable of establishing. Furthermore, probe membership information, such as DEG content, was integrated into the analysis/query space in various ways. These provided valuable contextual information which aided the user in the decision making process. The 6 probes identified earlier were of great interest to domain experts, due to the reason that will be discussed in Section 3.4. Again, such probes would be unlikely to be associated without the direct application of user knowledge and intuition.

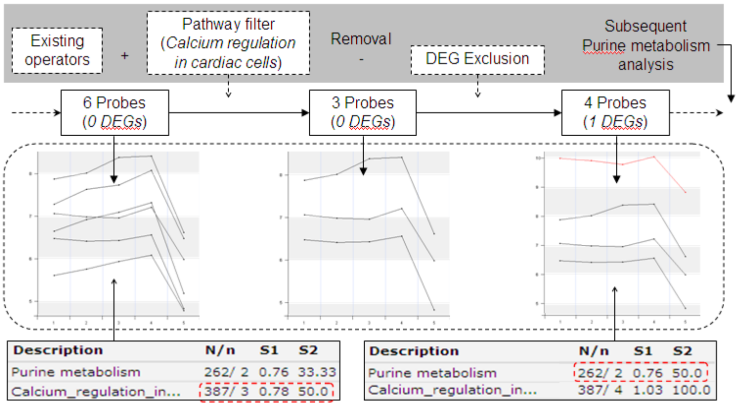
3.4 Interoperable Pathway Analysis

Biologists commonly want to identify the involvement of known biological processes in the observed time series. Systems such as GenMAPP, Ingenuity and GSEA provide mechanisms by which such pathways can be exposed, yet analysis within such systems is generally confined to DEGs. Statistical methods are employed to expose the most “significant” pathways represented, but issues relating to the completeness and quality of such subsets are here compounded.

Pathway analysis within XMAS can center on DEGs, as per traditional analysis, but is equally applicable to sets of probes sharing other characteristics – demonstrating the core value of interoperability. The power of XMAS to facilitate the exposition of probesets with significant commonality beyond, or in addition to differential expression, was explored in the previous scenario. Based on a serendipitous discovery, this scenario is extended, illustrating pathway analysis functionality within XMAS. This process is illustrated in Fig. 7.

It was indicated by the pathway membership view accompanying the visualization of the six probed from the previous use case (see Fig. 7), that the set has a significant three probe overlap with the pathway of *Calcium regulation in cardiac cells*. Interestingness measures provided quantitative support for the discovery, and the application of a corresponding pathway filter concentrated analysis on the three matching probes. DEG probes were reintroduced into the analysis space, revealing a single DEG sharing the developed characteristics. The appropriateness of the association of the additional DEG with the existing three probes was confirmed visually, and with the aid of the correlation matrices.

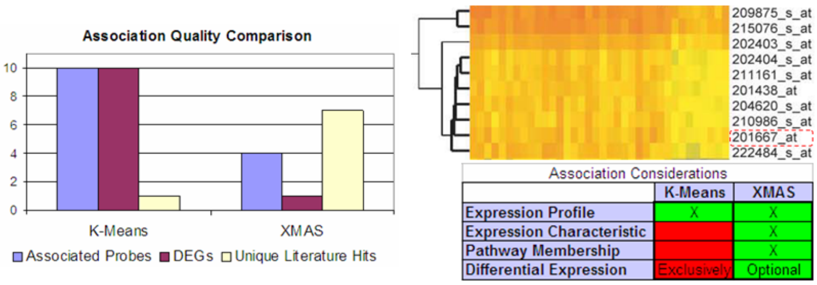
Serendipitous Discoveries: The exposure of 6 non-DEG probes, with a shared trajectory characteristic and expression profile led to the analysis of a pathway, which was unlikely to be judged significant by traditional analysis that focuses entirely and globally on the set of DEGs. The workflow that led to the association of non-DEGs with DEGs provided evidence to suggest that the localized observation was significant. Domain experts agree that the finding is striking, strengthening its candidacy for web lab experimentation. Further from the analysis of *Calcium Regulation*, the user noted a pathway overlap with *Purine metabolism*. This provides another extension point to analysis, which could manifest as a reverse analysis from local observation to global view of the relative involvement of *Purine metabolism*. *Smooth muscle contraction* is another such extension point.



Description	N/n	S1	S2
Purine metabolism	262/ 2	0.76	33.33
Calcium_regulation_in...	387/ 3	0.78	50.0

Description	N/n	S1	S2
Purine metabolism	262/ 2	0.76	50.0
Calcium_regulation_in...	387/ 4	1.03	100.0

Fig. 7. Workflow for the exploration of a serendipitous pathway discovery



Association Considerations

	K-Means	XMAS
Expression Profile	🟩 X	🟩 X
Expression Characteristic	🟥	🟩 X
Pathway Membership	🟥	🟩 X
Differential Expression	🟥 Exclusively	🟩 Optional

Fig. 8. Comparative assessment of association quality: The left figure compares the two sets of probes associated by k-means and XMAS respectively; the upper right figure identifies the common probe shared by these two sets; and the lower right table compares the factors under consideration by K-means and XMAS

Traditional analysis and analysis within XMAS are difficult to compare directly because of the differing emphasis on interaction and exploratory analysis, and global statistical/algorithmic analysis respectively. The outputs from both traditional and experiential approaches are comparable, however.

K-means analysis from 0, for example, associated the DEG from our set of four (201667_at) with 9 other DEGs, based on expression alone. This set serves as a direct comparison for the set of four which emerged from the previously described analysis. Despite having more probes, and more DEGs, the literature hits for our set far outweigh the expression (only) based association of k-means. See Fig. 8 for details.

4 Conclusions

This paper has presented XMAS, a web application developed with a new design philosophy to foster increased human-computer synergy. Various interoperable operators have been presented which combine with visualizations and HCI to compose an exploratory, interactive analysis system. Detailed use cases and comparisons made between XMAS and well established microarray analysis methods present evidence to prove the ability of this new approach to dramatically enhance the users experience during analysis. This materializes in the form of new, more complete hypothesis generation.

References

[1] Butte, A.: The use and analysis of microarray data. Nat. Rev. Drug. Discovery 1(12), 951–960 (2002)

[2] Microarrays: Chipping away at the mysteries of science and medicine, NCBI Just the Facts Series, `http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html`

[3] Faramarz, V.: Pattern Recognition Techniques in Microarray Data Analysis: A Survey. Bioinformatics and Medical Informatics (980), 41–64 (2002)

[4] Aas, K.: Microarray Data Mining: A Survey. NR Note SAMBA/02/01 (2001)

[5] Mukherjee J.: ICB, http://chagall.med.cornell.edu/I2MT/MA-tools.pdf

[6] Ingenuity Systems, http://www.ingenuity.com

[7] Draghici, S., et al.: Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. NAR 31(13), 3775–3781 (2003)

[8] Salomonis, et al.: GenMAPP 2: new features and resources for pathway analysis. BMC Bioinformatics 8, 217 (2007)

[9] Gentleman, R.C.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biology (2004)

[10] Project R, http://www.r-project.org/

[11] Tusher, V., et al.: Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98, 5116–5121 (2001)

[12] Tibshirani, R., et al.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 99, 6567–6572 (2002)

[13] Liang, Y., Kelemen, A.: Associating phenotypes with molecular events: recent statistical advances and challenges underpinning microarray experiments. Functional & Integrative Genomics 6(1) (January 2006)

[14] Singh, R., Jain, R.: From Information-Centric to Experiential Environments. In: Goldin, D., Smolka, S., Wegner, P. (eds.) Interactive Computation: The New Paradigm, pp. 323–351. Springer, Heidelberg (2006)

[15] Jain, R.: Experiential computing. Commun. ACM 46(7), 48–55 (2003)

[16] Kanehisa, M., et al.: KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480–D484 (2008)

[17] OmniViz, http://www.biowisdom.com/content/omniviz

[18] Lönstedt, I., Speed, T.P.: Replicated microarray data. Stat. Sin. 12, 31–46 (2002)

[19] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet. 25, 25–29 (2000)

[20] Winn, V., et al.: Gene Expression Profiling of the Human Maternal-Fetal Interface Reveals Dramatic Changes between Midgestation and Term. Endocrinology 148(3)