Expresso and Chips: Creating a Next Generation Microarray Experiment Management System

Allan Sioson*, Jonathan I. Watkinson†, Cecilia Vasquez-Robinet†, Margaret Ellis*,
Maulik Shukla*, Deept Kumar*, Naren Ramakrishnan*, Lenwood S. Heath*, Ruth Grene†,
Boris I. Chevone†, Karen Kafadar♯, and Layne T. Watson*
*Department of Computer Science
†Department of Plant Pathology, Physiology, and Weed Science
♯Department of Statistics
Virginia Tech, Blacksburg, VA 24061, USA

Abstract

Expresso is an experiment management system that is designed to assist biologists in planning, executing, and interpreting microarray experiments. It serves as a unifying framework to study data-driven application composition systems, as envisaged under the NSF Next Generation Software (NGS) program. Physical and analytical stages of the microarray process are mirrored in Expresso with computational models from biophysics, molecular biology, biochemistry, robotics, image processing, statistics, and knowledge representation. These models are pushed deeper (earlier) into the design process to help avoid costly design errors and to provide, as needed, surrogate functions for the traditional stages of microarray experiments. In this paper, we describe ongoing work in the design of Expresso, with specific reference to application composition, application optimization, experiment protocol design, and 'closing the loop.'

1. Introduction

Microarrays (sometimes referred to as *DNA chips*) have emerged as a promising approach to studying all the genes in a given organism simultaneously [12]; originally inspired by miniaturization trends in microelectronics, they have become an important technique in the bioinformatician's toolchest, and are constantly evolving to achieve higher value and to fit new uses. Applications using microarrays are characterized by the need to couple experimental design with data-driven analyses, to computationally model the many physical stages of the process, and to track and identify data and physical entities as they proceed through the laboratory and computational pipeline. They thus constitute a fertile ground for computer science research in application composition systems and dynamic data-driven application systems (DDDAS) [4], as envisioned by the CADSS (Complex Application Design and Support Systems) component of the National Science Foundation's Next Generation Software (NGS) program.

Our microarray experiment management system (*Expresso*) is designed with the above considerations in mind and aims to assist biologists in planning, executing, and interpreting microarray experiments. From an NGS perspective, Expresso is a data-driven application composition system that integrates models of varying fidelity to support, inform, and optimize the microarray design and analysis process. In this paper, we describe the motivations for Expresso, the architectural design of the system, and preliminary results from a biological study to elucidate the drought stress response of loblolly pine trees [14]. We assume basic knowledge of genomic analysis and bioinformatics terminology (e.g., see [9, 11]).

2. Microarray Technology

Two main approaches to preparing microarrays are based on complementary DNA (cDNA) technology and oligonucleotide synthesis. In the first, DNA templates (*probes*) are printed onto a high-density 2D array in a very small area on a solid surface. Typically the probes are chosen to correspond to all available genes that can be expressed in a given organism. The goal then is to determine the genes that are expressed when cells are exposed to experimental conditions, such as drought, stress, or toxic chemicals. To accomplish this, RNA molecules (*targets*) are extracted from the exposed cells and *reverse transcribed* to form complementary DNA (cDNA) molecules. These molecules are then allowed to bind (*hybridize*) with the probes on the mi-

croarray, which will adhere only with the locations on the array corresponding to their DNA templates. The cDNA target molecules are tagged with fluorescent dyes, so their binding to probes on the glass surface can be assessed by measuring the signal intensity using a laser. Intensity differences in spots will correspond to differential expression levels for particular genes. Using this approach, one can 'measure transcripts from thousands of genes in a single afternoon' [12].

In the oligonucleotide synthesis approach, a masking protocol synthesizes the probes *in situ* directly on the array, nucleotide by nucleotide. Oligo lengths are typically short (25–50 bases) which allows a higher packing density than with cDNA microarrays. The probe set constituting an oligonucleotide array involves selected sequences that identify particular genes. The design and manufacture of an oligonucleotide microarray is thus appropriate when a great deal of information is known about the expressed sequences in the studied organism. The array fabrication process is also very controlled and can be configured for high sensitivity, specificity, and repeatability. In contrast, cDNA microarrays have to contend with nonspecific hybridization, variation in number of molecules deposited per spot, and hybridization dependent on the length of the clone seuqence. When using cDNA arrays, the accepted practice then is to contrast the amount of hybridization observed to that experienced under 'control conditions.' For this reason, two types of mRNA target molecules are obtained — one from a *control* population and one from a *treated* population — and tagged with different dyes, that fluoresce at different frequencies. After hybridizing a mixture of these two populations onto the array (see Fig. 1), lasers of appropriate frequencies are used to read the signal intensity arising from the different populations — allowing the relative level of gene expression to be assessed.

The approach in Expresso is to mirror the physical and analytical stages of microarray experiment management using computational models from biophysics, molecular biology, biochemistry, robotics, image processing, statistics, and knowledge representation. These models are pushed deeper (earlier) into the design process to help avoid costly errors and to provide, as needed, surrogate functions for the traditional stages of microarray experiments. Our current emphasis is on cDNA microarrays, though some of the models extend and apply to oligonucleotide arrays. Before we motivate the usefulness of such modeling in the next section, it will be instructive to more carefully outline the sequence of steps involved in cDNA microarray preparation, experiment design, data gathering, and analysis (see Fig. 1), especially with an eye towards how errors are introduced and propagated. We discuss these steps under the categories of (a) probe generation and microarray design, (b) target preparation and hybridization, and (c) data generation and analysis. Some of these steps are specific to our current experimental process but the overall flow is indicative of almost all microarray experiments. Where appropriate, we identify sources of error, suggesting opportunities for modeling.

Probe Generation and Microarray Design

Clone Library Creation

cDNA clones (typically obtained from pools of messenger RNA) are housed in plasmid vectors, suitable for replication within a bacterial host. Extensive use of recombinant DNA technology facilitates the replication, after which the clones are extracted and archived as a plasmid cDNA library. These clones are then sequenced and annotated using tBLASTx similarity search [1]. The plasmid DNA is then housed in 96-well archive plates (8 rows by 12 columns) at -20°C. Sources of error at this stage include carryover of extraneous bacterial DNA into the library, missequencing, and mislabeling (e.g., incorrect annotation).

Clone Selection and Preparation

Given a set of clones to be studied in an experiment, the material is physically transferred from the clone library to working archive plates (of the same configuration). This is a manual process; sources of error are typically due to faulty material handling.

Dilution Transfers

The next step is to amplify the genetic material into quantities suitable for microarray experiments. Before this is effected, it is often important to dilute the working archive (100ng/μL) to usable concentrations (0.1ng/μL); typically one to four dilution transfers are made. Manual pipetting poses another source of material handling errors.

PCR Amplification

Plasmid DNA is transferred into a PCR plate (8 rows by 12 columns), which poses its attendant errors. *In vitro* techniques (as opposed to the endogenous bacterial machinery) amplify the quantity of the target cDNA. PCR stands for polymerase chain reaction and is a technique to amplify short stretches of DNA by specially designed 'primers' and the Taq polymerase enzyme (akin to a DNA Xerox). PCR errors involve amplifying portions of bacterial chromosomes (if present) and improper portions of the plasmid. The amplification efficiency is also related to the reaction conditions such as length of primer, annealing temperature, and number of cycles.

Cleaning Transfers

One to two cleaning transfers are made; this step uses vacuum wash stations to minimize carry over.

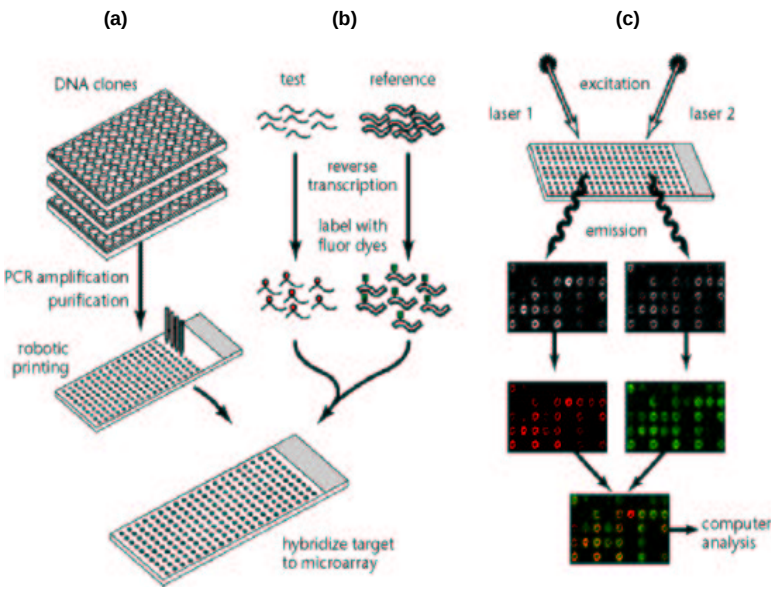

Figure 1. Schematic outline of a microarray design and analysis experiment. The three main steps are (a) probe generation and microarray design, (b) target preparation and hybridization, and (c) data generation and analysis. Figure courtesy J.M. Trent (National Institutes of Health) and reproduced with permission of Nature Genetics.

Transfer to Printing Plates
We now transfer the genetic material into 16×24 geometry plates (from four of the original 8×12 wells); this step involves a pipetting robot that does four rounds of 96 simultaneous transfers, providing as much as 10μL per well. An example of a pipetting robot is the TECAN GENESIS robot that can process multiple 96-well plates simultaneously. In the context of microarray experiments, the resulting printing plates provide for archival storage of cDNAs that go to the spotting robot (next step).

Transfer to Microarray
This is the last step of the array preparation process and uses a spotting robot such as the Affymetrix 417 arrayer; this robot features four pins that each takes material from different wells of a source microtitre plate and deposits one spot on a glass slide (typically 1nL/deposit). The geometrical configuration is often a transfer from six 16×24 plates to one 48×48 'subarray.' In our experiments, twelve such source plates contain enough material to create a hundred arrays (slides) with four 48x48 subarrays. In this and the previous step, detailed operations of the robots are available in the form of a programming language unique to the robot.

Target Preparation and Hybridization

RNA Isolation
The goal here is to isolate RNA from control and treated plant tissues. In this step, plant tissues are ground (powdered) under liquid nitrogen and placed in a buffer that maintains optimal conditions for RNA isolation (i.e., having the right pH, and being ribonuclease-free). The solution is extracted with chloroform and centrifuged (these two steps are repeated at least three times), in order to remove cellular debris, proteins, carbohydrates, and lipids. The protocols for these steps are available in [2]. We then precipitate the RNA overnight using a high salt solution (lithium chloride) at a low temperature ($-20°$C). After precipitation, we pellet the RNA with a centrifuge, rinse with an Ethanol solution, and resuspend in water. This typically produces 200–400 μgrams of RNA per sample.

Validation
We use spectrophotometric techniques and agarose gel analysis to ensure good RNA quality and quantity, primarily checking to see that it is intact and not degraded. In both this step and the previous one, all instruments, glassware, buffers, and solutions are treated in an autoclave to be RNase-free.

Reverse Transcription and Labeling
At this stage, the total RNA contains mRNA, tRNA and ribosomal RNA. Typically only 1-2% (!) is messenger RNA, the type important for hybridization. Reverse transcription occurs with *in vitro* techniques using MMLV (Moloney Murine Leukemia Virus) reverse transcriptase, random hexamer primers, free deoxyribonucleotides, and amino allyl labeled deoxyuridine, to produce complementary DNA (cDNA). The amino allyl is a side group that reacts with free amines. After reverse transcription Cy esters, which couple with the reactive amino allyl group, are added to the

cDNA. The cDNAs are thus labeled with fluorescent dyes. Control and treated samples are labeled with different dyes such that each sample will fluoresce at a different wavelength. This allows us to measure the differential level of gene expression after hybridization. Each comparison is often subjected to reciprocal labeling, to measure the effect of the dye.

Hybridization
We now pool the labeled control and treated samples, dry them down, and resuspend in a hybridization buffer. Essentially this ensures that conditions are right for hybridization; nonlabeled RNA and DNA is used to prevent nonspecific binding of target cDNA to the glass slide. The samples in the hybridization buffer are applied to the arrays, covered with a cover slip (another glass slide), sealed in chambers, and placed in a waterbath at 42°C overnight.

Washing
On the next day, arrays are washed in a series of post-hybridization wash buffers to remove nonspecific hybridizations.

Data Generation and Analysis

Image Generation
The slides are now ready to be scanned for assessing gene expression. Two lasers, one specific for each dye, excite the dye and the resulting fluorescence is measured using a scanner. The images are cataloged and the two images from a slide are superimposed for further analysis (see Fig. 1). Intensity levels from the two frequencies are processed to make a determination of differential gene expression (typically a ratio [3] or a log ratio).

Statistical Analysis
Due to the multitude of factors that can (and do) play a role in differential gene expression, the assessments made above must be substantiated by statistical analysis. The numerous models available at this stage attempt to quantify measurement error [22], correct for background noise [10], account for different types of interactions [31], assess any systematic variation across different regions of the array [15] and, more generally, aim to improve the robustness of gene expression estimates. Gene replicates and control genes for capturing the contributions due to many of these effects are often used in microarray experiments.

Data Mining
Since a given microarray experiment produces the expression level of hundreds or thousands of genes, data mining techniques such as clustering [6] and inductive logic programming [14] are employed for summarization and descriptive characterization of the results. Often prior biological knowledge is used to reconstruct metabolic networks [5] underlying the processes that are being studied.

3. Model-Based Design and Management of Experiments

In contrast to the variety of software tools for cataloging and analyzing microarray data, Expresso is designed to support modeling and control of the entire microarray experimental process. Model libraries in Expresso are meant for *application composition* (creating a model of multiple stages from models of individual stages), *application optimization* (configuring a process to satisfy desired performance constraints on model outputs), *experiment protocol design* (e.g., running a set of virtual experiments using the incorporated models to gather performance metrics about the ultimate set of real experiments, and using these metrics to refine the experimental protocol), and *closing the loop* (using information from later stages to better inform and suggest configurations for earlier stages in later experiments). All of these capabilities are becoming integral aspects of bioinformatics software systems [13]; in our discussion below, we present different stages of the microarray process that can benefit from such facilities.

3.1 Application Composition and Optimization

Successful microarray system design and execution requires mathematical modeling, and models at various levels of fidelity are available (e.g., for material selection, PCR amplification [27], spotting, hybridization [21], and image processing). However, due to the relative paucity of information available to drive these models and the disconnect between model builders (who study the behavior of biological models and understand the details of instrumentation) and biologists (who assemble the arrays), microarray design and management is still an imperfectly understood process.

The typical approach to designing a process (experiment) comprised of a sequence of operations is to independently optimize each operation, and then modify these process component optimal solutions to achieve compatibility between consecutive components of the process. For example, quality control has been investigated primarily from this perspective for the chemical synthesis of target molecules and PCR amplification (in the context of the entire microarray design process). The implicit belief is that the overall process design achieved in this way cannot be far from optimal, since each of its constituent components is (nearly) optimal. Unfortunately, this is rarely true, and the more complex the process the more likely this component-wise optimal design is to be far from a global optimum for the overall process. This fundamental fact is well known, and has led to engineering design methodologies known over the years as system engineering, integrated design, and most recently, multidisciplinary design optimization.

Historically, because of the computational cost, integrated systems design has used low fidelity, relatively cheap models for all the components of a process or system. After an approximately optimal global system design has been found, the individual components are then reoptimized using expensive, high fidelity models, subject to the constraints imposed by the overall global system design. Sometimes this process fails – the accurately reoptimized process components are found to be incompatible with the imposed constraints – owing to the inaccuracies of the low fidelity models used in the global system design. Thus, in practice, integrated design becomes an iteration, alternating between global system optimization involving low fidelity models and component optimization involving high fidelity models.

While the stringency of design requirements for high throughput microarray systems is acknowledged, the use of mathematical models to optimize the experimental process (including analysis) is not widespread (and sometimes not even acknowledged). Challenging algorithmic research questions include the management of models of varying fidelity, the effective combination of global with local optimization, and building functional approximations to sparse data in high dimensions (*surrogates* for the data, so to speak). The latter is especially important, since functional relationships, expressed mathematically, are more useful in many contexts than the raw data.

Our operative principle is that the cost of correcting design errors is directly related to how early they occur in the design cycle. The earlier decisions are made, the more expensive they are to modify later. The goal is then clearly to move high fidelity models and analysis earlier in the design cycle, or in the context of microarray experiments, deeper into the design. For instance, using a sophisticated titration model or selecting sequences based on a molecular dynamics energy minimization (e.g., as recently proposed by Lafontaine and Lavery [17] and by Shchyolkina et al. [23]) is likely to be more effective than compensating for bad microarray output data, or redoing the experiment with different sequences. The challenge is not only to move high fidelity modeling deeper into the experimental design process, but to judiciously combine data of varying fidelity, balancing accuracy requirements with computational time constraints.

One area that can benefit from such an approach involves modeling the PCR amplification of targeted DNA sequences. All ingredients necessary for DNA duplication are placed together in a vial; different temperatures are then successively applied to first separate the DNA strands, cool them down (so that a primer can bind), and then use a polymerase enzyme to extend the primer and create a complementary copy of the DNA strand. All these steps can be completed in at most two minutes; the sequence is actually then repeated multiple times to achieve high levels of replication. If we assume that the extension process is Markovian and occurs independently on each primer (e.g., see [27]), we can derive a mathematical expression for the distribution of length of the growing strand as a function of reaction conditions (temperature, rate coefficients, number of cycles, and the activation energy for dNTP addition).

Since amplification factors in excess of a million can be achieved, PCR is an extremely sensitive technology and the process must be carefully modeled to control the kinetics of the reaction. For instance, Velikanov and Kapral [27] describe how to capture aspects such as the 'flattening' of the yield of the target sequence with increasing number of cycles and sensitivity to the initial reaction conditions. Numerical optimization is hence necessary to determine an optimal schedule (cycle durations) for PCR, given factors such as the template length, primer length, and binding site offsets.

Beyond suggesting suitable duration sequences, such an optimized model can actually play a larger part in Expresso, namely, to serve as a surrogate model for gene quantification and help validate gene expression inferred at the end of the analysis pipeline (e.g., in the sense of quantitative, real-time PCR [18]). Here the approach is to use the PCR reaction dynamics to arrive at an estimate of the initial state of the system, namely the amount of nucleic acids.

3.2 Experimental Protocol Design

Tuning and optimizing experimental protocols is an important endeavor in laboratory-based biology. Expresso's design is intended to support the refinement of experimental protocols by using the incorporated models to conduct *virtual experiments* and gather important performance metrics. We outline this idea in the context of the hybridization step.

There are many factors that affect the performance of hybridization, where fluorescently labeled targets are brought together with immobilized DNA probes for a limited time under thermally controlled conditions. A key factor is whether oligonucleotide or cDNA microarray technology is employed [16]. Each DNA probe (oligo) on an oligonucleotide microarray is manufactured to have a specific nucleotide sequence and has a typical length of about 25 nucleotides (e.g., in Affymetrix chips http://www.affymetrix.com/technology/design/index.affx). On the other hand, each DNA probe (clone) on a cDNA microarray is an actual expressed sequence selected from a library of clones and amplified via PCR. A clone is hundreds of nucleotides long, and typically only a portion of its sequence is known. The lengths of clones on a cDNA microarray vary greatly, as do the lengths of expressed sequences (mRNA) among the targets.

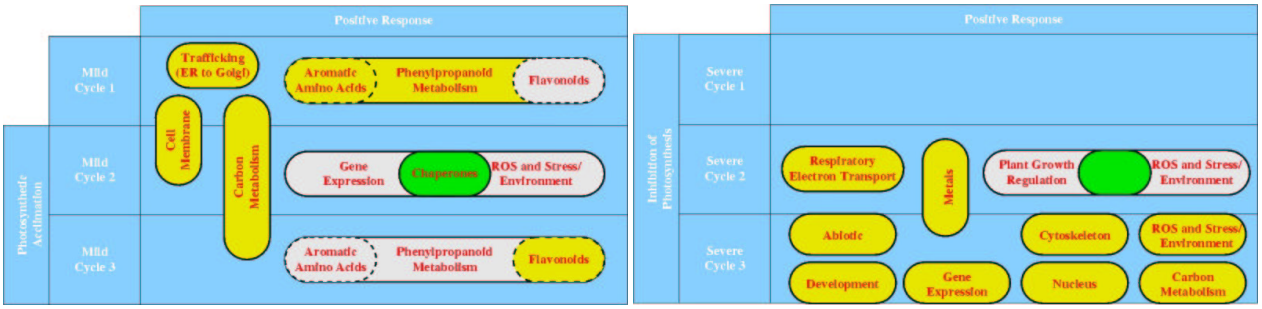

Figure 2. Functional categories mined by inductive logic programming corresponding to genes exhibiting (left) positive expression in mild stress condition (characterized by photosynthetic acclimation) and (right) positive expression in severe stress condition (characterized by inhibition of photosynthesis).

Other factors affecting hybridization include probe attachment method, probe purity, probe and target lengths, probe and target sequences, and concentrations. Relevant physical characteristics of the hybridization include the electrostatic potential and thermodynamic equilibrium [26]. The general phenomenon of hybridization of strands of nucleic acids is a longstudied process in biochemistry and related fields. The technique of subtractive hybridization, whose purpose is to isolate and amplify sequences expressed in small quantities, is more recent and has yielded some additional insights and models [8]. The phenomenon of hybridization in microarrays is different in that one of the strands (the probe) is immobilized by attachment to a surface; this leads to other empirical and theoretical results [20]. A particular phenomenon of critical concern for the analysis of microarray results is cross-hybridization, where a target with a sequence that is a slight mismatch to the sequence of a given probe may hybridize to the probe anyway [7].

Of special interest in Expresso are existing empirical, mathematical, simulation, and statistical models of hybridization that can be adapted to our modeling of hybridization in oligonucleotide or cDNA microarrays. Tu, Stolovitzky, and Klein [25] have analyzed factors that influence noise in the results of microarray experiments and have obtained equations for noise distributions. Vainrub and Pettitt [26] have developed a physical model of electrostatic forces that explains observations on the efficiency of hybridization between immobilized and free DNA as a function of probe surface density. Gadgil, *et al.*, [8] develop a mathematical model for the kinetics of subtractive hybridization that includes relevant factors such as strand length and hybridization temperature. Peterson, Wolf, and Georgiadis [21] study the kinetics of hybridization with an immobilized probe and suggest the use of one of two models, one for perfectly matched sequences and a second for mismatched sequences. Walton, *et al.*, [29] apply a mathematical model to the prediction of the thermodynamics and kinetics of hybridization and compare the predictions to experimental results. Combining models such as these can predict the range of results and noise that should be obtained from a particular microarray experiment before it is implemented. An examination of the predictions can reveal deficiencies in the characteristics of the experimental design. The predicted deficiencies can be used to refine the experimental protocol computationally, and hence produce an improved protocol, without the need for chemical reagents.

3.3 Closing the Loop

An important hallmark of Expresso is its support for closing-the-loop in microarray experiments. One complete cycle of all the stages in a microarray experiment results in both qualitative and quantitative assessments of gene expression, and these results can be used to design the next cycle of microarray experiments. For instance, Fig. 2 shows the results of a data mining algorithm (inductive logic programming [19]) applied to gene expression data from loblolly pine clones exposed to two different stress levels (mild and severe). The purpose of this data mining step is to *redescribe* clusters of gene expression values into functional categories (signifying gene membership). The two stress levels studied are physiologically characterized by photosynthetic acclimation and inhibition, respectively. Hence summarizing the data in terms of functional categories reveals insight into the nature of the different pathways involved. The goal now is to use these results to influence the design of the future experiments, by configuring the probe set to reflect the genes of interest.

This probe selection problem [24] can be posed as a discrete optimization problem having constraints such as minimizing cross-hybridization, accounting for alternative splicing sites, and increasing the specificity of results (as compared to the previous cycle of experiments). More generally, there are conflicting objectives in terms of increasing the coverage of the library and also deliberately introducing redundancy to account for the many experimental errors that occur in the microarray process.

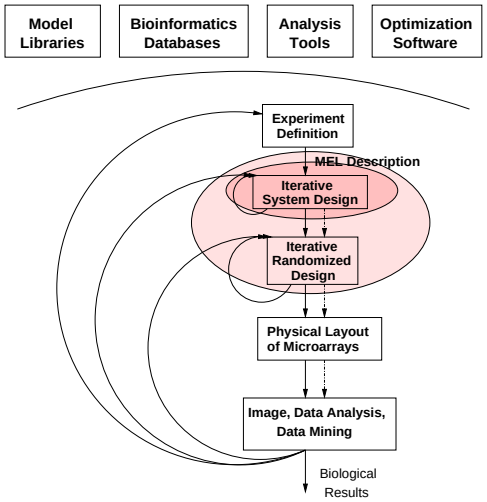

Figure 3. Execution and flows in Expresso. The solid lines indicate computational flows; the dashed lines indicate physical and material flows. The shaded regions highlight opportunities for automated optimization of composed models, currently being pursued.

4. Software Design of Expresso

Since the Expresso design underscores the importance of modeling both physical and computational flows, it is important that the software architecture provide for a constantly changing scenario (in terms of data, schema, and the nature of experiments conducted). The ability to provide expressive and high performance access to objects and streams (for experiment management) with minimal overhead (in terms of traditional database functionality such as transaction processing and integrity maintenance) is thus paramount in Expresso [28].

Fig. 3 describes the characteristics of physical and computational flows in Expresso. Programs written in our in-house microarray experiment markup language (MEL) direct the automatic synthesis of fragments from a model library and the creation of suitable database tables. There is a one-to-one mapping between MEL's specifications and the basic entities in the database schema. MEL supports basic validation, although in a form less restrictive to languages like XML Schema. For instance, it does not support referential integrity constraints or type derivation by extension – features that are heavy bottlenecks for memory consumption and speed of access.

One MEL description is typically employed for each individual microarray experiment. As shown in Fig. 3, multiple iterations involving system design and randomized layout design (designing placement of cDNAs in microtitre plates, and ultimately on slides) might be required before consigning the description to a physical experiment. These early stages can thus be configured to run as a set of virtual experiments using the incorporated models to obtain an indication of the complexity of the ultimate set of real experiments. When results of the desired fidelity are achieved, the biologist can choose to physically realize the microarray experiment and use any collected data to improve or optimize any aspect of the modeling (ranging from reorganizing the categorical assignment of clones to updating the randomized layouts).

One design feature of Expresso that aids in such reactive execution of compositional modeling steps is the use of active database elements such as triggers and rule systems. Triggers are already accepted in scheduling and workflow management where they perform functions ranging from integrity maintenance to audit trails for system administration [30]. In Expresso, we are investigating an intermediate approach between fully automatic trigger generation and completely handcrafted trigger sets. The declarative specifications of MEL provide one starting point for encoding triggers and can thus model strategies for closing the loop.

5. Discussion and Ongoing Work

Two cycles of preliminary microarray experiments are complete with *Pinus taeda* (loblolly pine) as our model plant for stress studies (see, for instance [14]). We are now in the process of populating our model library with many relevant codes and models (both descriptive and predictive), which will then be used for optimizing various aspects of the microarray process. In the coming year, we will adapt Expresso functionality to study systems based on *Picea abies* (Norway spruce), *Solanum Tuberosum* (potato), and *Homo Sapiens*, besides loblolly pine.

One of our long-term goals is to use Expresso to design multipurpose microarrays; current use of microarrays is almost entirely restricted to probes and targets from the same species. By using sequences that are conserved among genes with identical function in different species, it may be possible to construct a *heterologous chip*, a microarray of probes that hybridize to cDNAs obtained from closely related species. Thus one microarray design would be capable of yielding useful information about gene expression in numerous species. Such a multipurpose microarray would be ideal for studies on grass species of agronomic importance (e.g., rice, tall fescue, Kentucky bluegrass), only some of which have been completely sequenced. To design arrays of such broad functionality, it is imperative that model-based design be employed, to predict and compare the effects of various sequences on different aspects of the microarray process. Such computational capabilities do not exist presently; Expresso's approach to microarray experiment management promises to support such modes of inquiry.

Acknowledgements

This work is supported by US NSF grant EIA-0103660 entited 'A Microarray Experiment Management System.'

References

[1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, Vol. 215:pages 403–410, 1990.

[2] S. Chang, J. Puryear, and J. Cairney. A Simple and Efficient Method for Isolating RNA from Pine Trees. *Plant Molecular Biology Reporter*, Vol. 11:pages 113–116, 1993.

[3] Y. Chen, E. Dougerty, and M. Bittner. Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. *Journal of Biomedical Optics*, Vol. 2:pages 367–374, 1997.

[4] F. Darema. Dynamic Data-Driven Application Systems. In D. Marinescu and C. Lee, editors, *Process Coordination and Ubiquitous Computing*. CRC Press, 2002.

[5] J. DeRisi, V. Iyer, and P. Brown. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*, Vol. 278:pages 680–686, 1997.

[6] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proceedings of the National Academy of Sciences, USA*, Vol. 95:pages 14863–14868, 1998.

[7] E. Evertsz, J. Au-Young, M. Ruvolo, A. Lim, and M. Reynolds. Hybridization Cross-Reactivity within Homologous Gene Families on Glass cDNA Microarrays. *Biotechniques*, Vol. 31(5):pages 1182,1184,1186, November 2001.

[8] C. Gadgil, A. Rink, C. Beattie, and W. Hu. A Mathematical Model for Suppression Subtractive Hybridization. *Comparative and Functional Genomics*, Vol. 3(5):pages 405–422, Oct 2002.

[9] G. Gibson and S. Muse. *A Primer of Genome Science*. Sinauer Associates, Inc., Dec 2001.

[10] A. Goryachev, P. Macgregor, and A. Edwards. Unfolding of Microarray Data. *Journal of Computational Biology*, Vol. 8(4):pages 443–461, 2001.

[11] A. Griffiths, J. Miller, D. Suzuki, R. Lewontin, and W. Gelbart. *Introduction to Genetic Analysis*. W.H. Freeman and Company, New York, 1999.

[12] H. Hamadeh and C. Afshari. Gene Chips and Functional Genomics. *American Scientist*, Vol. 88:pages 508–515, Nov-Dec 2000.

[13] L. Heath and N. Ramakrishnan. The Emerging Landscape of Bioinformatics Software Systems. *IEEE Computer*, Vol. 35(7):pages 41–45, July 2002.

[14] L. Heath, N. Ramakrishnan, R. Sederoff, R. Whetten, B. Chevone, C. Struble, V. Jouenne, D. Chen, L. van Zyl, and R. Grene. Studying the Functional Genomics of Stress Responses in Loblolly Pine with the Expresso Microarray Experiment Management System. *Comparative and Functional Genomics*, Vol. 3(3):pages 226–243, June 2002.

[15] M. Kerr, M. Martin, and G. Churchill. Analysis of Variance for Gene Expression Microarray Data. *Journal of Computational Biology*, Vol. 7(6):pages 819–837, 2000.

[16] W. Kuo, T. Jenssen, A. Butte, L. Ohno-Machado, and I. Kohane. Analysis of Matched mRNA measurements from Two Different Microarray Technologies. *Bioinformatics*, Vol. 18(3):pages 405–412, March 2002.

[17] I. Lafontaine and R. Lavery. Optimization of Nucleic Acid Sequences. *Biophysical Journal*, Vol. 79:pages 680–685, August 2000.

[18] J. Larrick (Ed.). *The PCR Technique: Quantitative PCR*. Eaton Publishers, June 1997.

[19] S. Muggleton. Scientific Knowledge Discovery using Inductive Logic Programming. *Communications of the ACM*, Vol. 42(11):pages 42–46, November 1999.

[20] A. Peterson, R. Heaton, and R. Georgiadis. The Effect of Surface Probe Density on DNA Hybridization. *Nucleic Acids Research*, Vol. 29(24):pages 5163–5168, Dec 2001.

[21] A. Peterson, L. Wolf, and R. Georgiadis. Hybridization of Mismatched or Partially Matched DNA at Surfaces. *Journal of the American Chemical Society*, Vol. 124(49):pages 14601–14607, Dec 2002.

[22] D. Rocke and B. Durbin. A Model for Measurement Error for Gene Expression Arrays. *Journal of Computational Biology*, Vol. 8(6):pages 557–569, 2001.

[23] A. Shchyolkina, O. Borisova, M. Livshits, G. Pozmogova, B. Chernov, R. Klement, and T. Jovin. Parallel-Stranded DNA with Mixed AT/GC Composition: Role of Trans G-C Base Pairs in Sequence Dependent Helical Stability. *Biochemistry*, Vol. 39:pages 10034–10044, 2000.

[24] S. Tomiuk and K. Hofmann. Microarray Probe Selection Strategies. *Briefings in Bioinformatics*, Vol. 2(4):pages 329–340, Dec 2001.

[25] Y. Tu, G. Stolovitzky, and U. Klein. Quantitative Noise Analysis for Gene Expression Microarray Experiments. *Proceedings of the National Academy of Sciences, USA*, Vol. 99(22):pages 14031–14036, Oct 2002.

[26] A. Vainrub and B. Pettitt. Coulomb Blockage of Hybridization in Two-Dimensional DNA Arrays. *Physical Review E.*, Vol. 66(4), Oct 2002.

[27] M. Velikanov and R. Kapral. Polymerase Chain Reaction: A Markov Process Approach. *Journal of Theoretical Biology*, Vol. 201:pages 239–249, 1999.

[28] A. Verstak, N. Ramakrishnan, L. Watson, J. He, C. Shaffer, K. Bae, J. Jiang, W. Tranter, and T. Rappaport. BSML: A Binding Schema Markup Language for Data Interchange in Problem Solving Environments. *Scientific Programming*, Vol. 11(1), 2003. to appear.

[29] S. Walton, G. Stephanopoulos, M. Yarmush, and C. Roth. Thermodynamic and Kinetic Characterization of Antisense Oligodeoxynucleotide Binding to a Structured mRNA. *Biophysical Journal*, Vol. 82(1):pages 366–377, Jan 2002.

[30] J. Widom and S. Ceri. *Active Database Systems: Triggers and Rules for Advanced Database Processing*. Morgan Kaufmann, 1996.

[31] R. Wolfinger, G. Gibson, E. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. Paules. Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models. *Journal of Computational Biology*, Vol. 8(6):pages 625–637, 2001.