GeneX: An Open Source gene expression database and integrated tool set

by H. Mangalam, J. Stewart, J. Zhou, K. Schlauch, M. Waugh, G. Chen, A. D. Farmer, G. Colello, J. W. Weller

Because gene expression profiles are highly sensitive to sample and processing conditions, it is crucial to accurately represent these conditions along with the numeric data in a way that allows the conditions to be part of a query. The GeneX™ project is intended to provide an Open Source database and integrated tool set that will allow researchers to store and evaluate their gene expression data and, moreover, such evaluation will be independent of the technology used to obtain the data.

The genomics revolution is beginning to produce data at a rate that will soon rival that of high-energy physics. Biologists and informaticists are faced with critical issues of how best to organize this information and share it so that it provides the greatest good to the greatest number of researchers.

The "genomic" data type, as produced by the various genome projects, is a linear DNA (deoxyribonucleic acid) polymer consisting of four basic nucleotides (A, C, G, T) repeated nonrandomly in strings of up to several million (the length of a chromosome).[1] In the human genome, there are approximately three billion bases distributed among 23 chromosomes, but despite the diversity observed in humans, this sequence is 99.99 percent invariant between individuals. Completion of the Human Genome Project[2] will reveal the precise order of the sequence, both where it is invariant and how it varies. This genomic sequence information encodes the range of responses that an organism can deploy to cope with its environment but is itself largely static:[3] the genome itself does not change, but the information derived from it does.

The number of genes in a genome ranges between 20, the number of genes in a virus, and 30000, the number of genes in a human. The average human gene is about 3000 bases in length, so the amount of gene sequence in the human genome that is used for coding genes is quite small, approximately 1 percent. Each gene (a contiguous section of DNA) can be *transcribed* to produce mRNA (messenger ribonucleic acid), which in turn is *translated* into a unique protein (a linear polymer of amino acids). The protein can subsequently be further modified by a variety of mechanisms, such as phosphorylation, glycosylation, and cleavage. Adding to this complexity, one gene can code for several unique mRNAs via a mechanism called splicing, in which parts of the sequence (called intervening sequences or introns) are spliced out and the remaining coding sequences (exons) can be "spliced" to form several unique mRNAs. Regulation of gene transcription occurs through a complex feedback mechanism involving a number of pathways, ultimately being mediated by transcription factors, protein complexes that bind to short regulatory sequences of DNA near the start of transcription.

In contrast to the static genomic DNA, gene expression is the dynamic response of the genome to environmental conditions. As such, gene expression data require more descriptive information (metadata) to characterize it accurately: the mRNA is ex-

0018-8670/01/$5.00 © 2001 IBM

pressed at different times and in varying amounts during an organism's life. In multicellular organisms, cells of a particular tissue express unique cohorts and profiles of genes in amounts appropriate to cell type. It is critical to understanding gene expression that variables, such as cell type and developmental stage as well as perturbations imposed by the researcher, be documented.

Measurement of gene expression involves determining the amount of a unique mRNA in a complex sample that is harvested at a particular time after response to a specific physiological condition. The techniques to measure quantities of mRNA are varied. In most high-throughput gene expression array technologies, known DNA sequences are placed on the supporting media as probes to quantify the hybridizing sequences, which are typically labeled with a radioisotope or one or more fluorescent dyes. Image analysis permits the raw signal and background noise to be parameterized in a number of ways (roundness of spot, profile of pixel intensities, signal to background, background-subtracted). Ultimately the level of gene expression is given as an intensity value per spot or area or pixel, based on an empirically optimized expression that seems to provide the best overall accuracy. The immobilized DNA can range in size from only a few bases (an oligonucleotide) to several thousand bases, which may represent a complete mRNA. This technique is highly scalable, with the result that today, tens of thousands of genes can be measured simultaneously.

This area of research is of interest not only for understanding the basic biology of gene expression, but also for its applicability to clinical medicine. As the technology improves and the price drops, there is the promise of personalized medical treatment based on a much more accurate diagnosis of an ailment. Identification of the exact effects of HIV infection on an individual's expression profile,[4] differentiation between cancers having apparently identical pathologies to identify specific treatment or sensitivity,[5,6] and identification of a patient's possible exceptional resistance or weakness to a particular treatment[7] have been postulated as reasonable outcomes of widespread use of this technology.

The large volume of data produced by gene chip or array experiments has dictated the use of computerized data structures to store the data. For some labs this is no more than flat files and spreadsheets, but such simple storage mechanisms are extremely inefficient for retrieval of particular subsets of the results. Several more advanced gene expression databases exist: most focus on either a particular technology or a particular organism or both. While there are some interesting commercial gene expression databases, produced by Rosetta Inpharmatics[8] and GeneLogic,[9] the specifics of their internal structure are not available for public scrutiny due to their proprietary nature. Some noncommercial efforts to design more general databases merit particular mention; three of the most promising are described next.

Array Express[10] is a data model produced by the European Bioinformatics Institute. Some minor changes to it allowed it to be instantiated as a relational database, and MaxdSQL[11] is the resulting implementation. The principal difference between GeneX** and ArrayExpress is in the definition of the atomic data unit. In MaxdSQL, the *Experiment* is the basic data unit and it is not possible to create virtual experiments from subsets of arrays. The *Gene* class has a name that is given by the submitter without reference to controlled vocabularies, causing a proliferation of names for identical genes. The types of manipulations are limited to time and temperature series, although this restriction can easily be relaxed. The authors have done an excellent job of coding the interfaces, not only for MaxdSQL, but also its sister applications MaxdLoad (a data-loading application somewhat like our Curation Tool—see later in this paper) and MaxdView (an analysis application). Work is under way to make MaxdSQL compatible with GeneX.

The Gene Expression Omnibus[12] (GEO) is a gene expression database hosted at the National Library of Medicine. It supports four basic data elements: Platform (the physical reagents used to generate the data), Sample (information about the mRNA being used), Submitter (the person and organization submitting the data), and Series (the relationships among the samples). Further, while it allows the download of entire data sets, it has no ability to query the relationships via SQL (Structured Query Language) or other means. Data are entered as tab-delimited ASCII (American National Standard Code for Information Interchange) records, with a number of columns that depend upon the type of array that has been selected as the data source. It does support Serial Analysis of Gene Expression[13] (SAGE) data and there has been an effort to incorporate several levels of analyzed data, much like the "derived data types" described later.

The Stanford MicroArray Database[14] (SMD) is one of the first academic databases to be used on an institutional scale. It contains the largest amount of data of any academic database, due to its close association with one of the first groups to develop large-scale arrays. It is similar to GeneX in that it uses a relational database to answer queries. Some of its limitations are: it supports only Cy3/Cy5 glass slide data, it is designed to exclusively use an Oracle database, and it has not been released outside Stanford.

GeneX differs from the databases above in two ways.

1. It represents more relevant biological data. Not only does it support multiple species and gene expression technologies, but it also specifically addresses the dichotomy between the immobilized nucleic acid sequences that are used as hybridization targets and the genes that they are meant to represent.

2. It was designed from the start to be installed in multiple sites, thus creating a peer-to-peer federation of databases. Moreover, the communication of gene expression data is through the use of an XML-based (Extensible Markup Language) language, GeneXML, described later.

The rest of this paper is organized as follows. First, we describe the system design and give an overview of the main components. We then describe the data model and its support for both meta-data and numeric data. Next, we discuss the technologies used to overcome the challenge of moving the gene expression data from the laboratory and into the database: the client-side Curation Tool, the GeneXML for encoding the data, and Genex.pm, the Perl wrapper that provides the application programming interface (API) for manipulating the database. We then describe the user interface for the query and the analysis tools, and conclude with a summary of the results presented here.

System design

Unlike many business environments, there is significant platform heterogeneity in the biology community. Microsoft Windows** is the most popular platform, but Apple's Macintosh** is used by a large minority, and there is a small but growing number of UNIX** and Linux** platforms, especially in the bioinformatics and computational biology areas. Therefore GeneX client software was designed to be as platform-neutral as possible. The GeneX system is specifically designed to incorporate Open Source Software (OSS). For example, an OSS application from AT&T called Virtual Network Computing[15] (VNC) allows the X Window System** and Microsoft Windows and Apple Macintosh systems to provide and receive displays from each other. GeneX uses HyperText Markup Language (HTML) forms as the preferred simplest interface for both input and output. Where more interactivity benefits users, as with the Curation Tool (a Java** application), a more complex interface is provided.

We chose the Linux operating system running on Intel hardware as the most cost-effective server platform for the widest audience. We also support the Solaris/SPARC** platform, but anticipate that most installations, especially first-time users, will be running Linux.

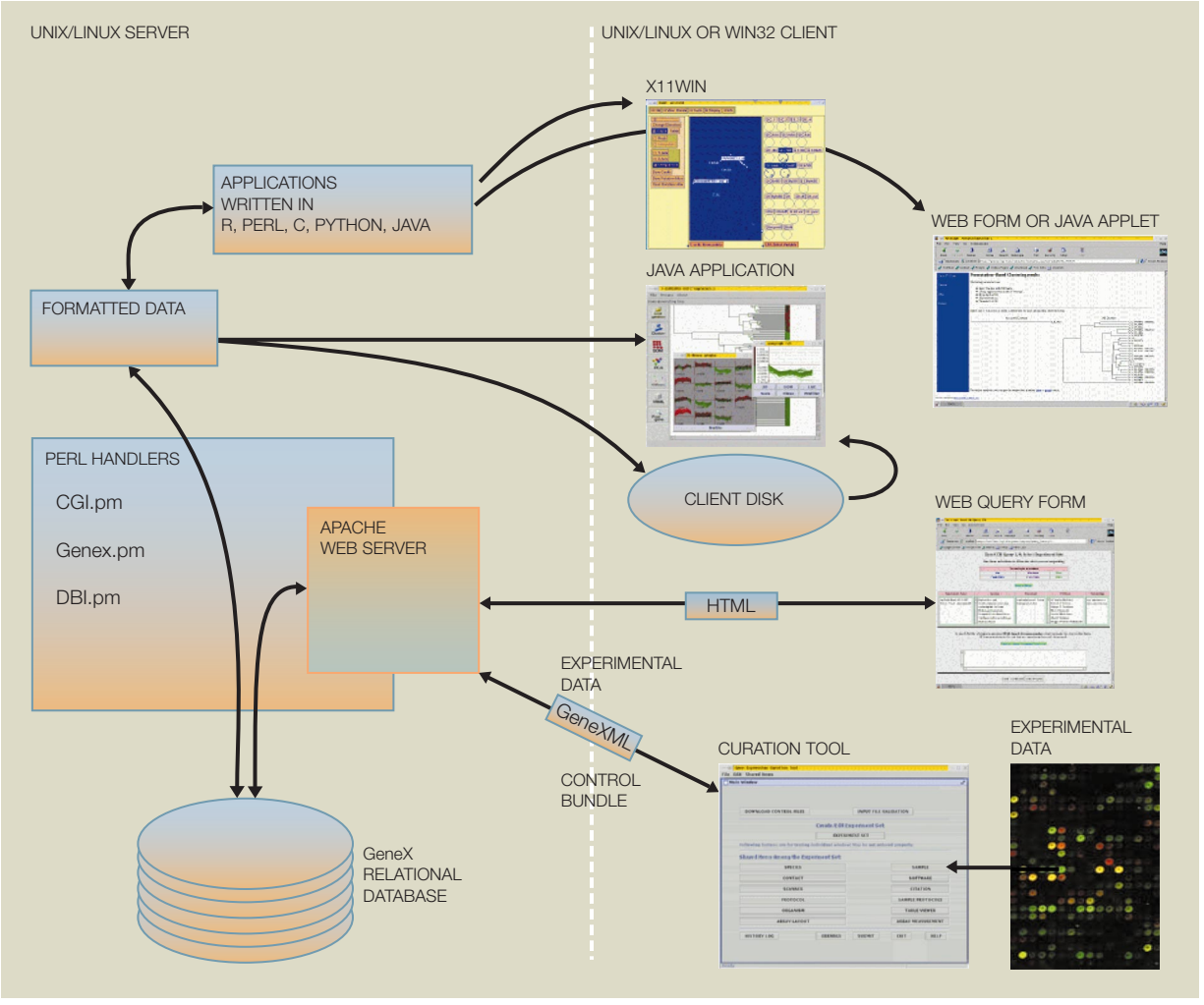
The GeneX package has been assembled as an Open Source project, released under the Free Software Foundation's *Lesser General Public License*.[16] This protects the core of GeneX from commercial forking, but allows it to be used with proprietary replacement components and add-ons. The system is freely available to anyone in source code form via anonymous CVS (Concurrent Versioning System) checkout from SourceForge.[17]

Components. As illustrated in Figure 1, GeneX is a Web-based design that can be implemented either as an entry-level system (Linux/Apache/PostgreSQL) or a high-end system (Solaris**/Netscape**/Sybase**). It uses HTML forms for most of the input and output but also some Java-based applets, and some X Window System applications that require either an X server or the free VNC system. It also uses GeneXML, an XML[18] language.

Also required to run GeneX are the Apache[19] Web server, Perl,[20,21] the statistical modeling language R,[22] Java,[23] and PostgreSQL[24,25] relational database system. Optional components include J-Express,[26] Stanford's xcluster,[27] and TreeView[28] (the last two are free of charge to academic and nonprofit organizations).

As seen in Figure 1, data flows from the Java Curation Tool (lower right) via GeneXML encrypted by secure sockets layer (SSL) to the Web server, where it is placed in a queue for Curator approval. Once approved, it is uploaded into the database via a set of Perl scripts included in the Genex.pm module.

Figure 1 Structural diagram of the GeneX prototype



Users query the database by filling out HTML forms which are processed by Perl scripts in CGI.pm.

The query results can be downloaded in several formats (tab-delimited text, GeneXML, etc.) or can be passed directly to analytical routines written in the R statistical language. Most of these analytical routines produce either static images or alphanumeric data, which can be displayed in HTML. Other output can be formatted for display using sophisticated three-dimensional plotting and built-in data manipulation routines (see later).

To facilitate data warehousing, the GeneX database is based on an abstract model of gene expression data. Results of experiments that can be represented by sequence data, identifiers, and signal measurements can be stored in the database. The basic quantitative data are associated with related information concerning biological and environmental parameters. Capturing and organizing these "meta-data" establishes a basis for an ontology for gene expression experiments. Some of the more important relationships between objects in the hierarchical system are described in detail below.

Security is established by recognizing the data providers and their designated partners as owners of the experimental data. Access to the database is pass-

word-protected and data providers may only update their own data.

Gene expression data are structured and suitable for representation by XML. The NCGR (National Center for Genome Resources) has released a document type definition (DTD) for a gene expression markup language (GeneXML) and NCGR is also part of the gene expression consortium that is developing a community-wide XML standard.

The primary expression data must be represented at a sufficient level of abstraction to make the data independent of the source technology.

The uploading of large, disparate sets of data to a database is a potential minefield of errors. The Curation Tool is an integral part of GeneX designed to address this problem. It downloads species-appropriate, controlled vocabularies of taxonomies, protocols, gene names, and other details from the selected database to ease the documentation of new data sets by populating forms with the appropriate information. The tool associates the information selected in this way with the numeric data arising from a particular experiment (e.g., the quantities of each gene and variations in experimental parameter values). Those fields requiring certain types of data in specific formats are validated before upload to a database, allowing the owner of the experiment to correct errors before they reach the database. Data processed by the Curation Tool is uploaded to a GeneX database formatted in GeneXML. Data from other GeneX databases may be downloaded and read by the Curation Tool as GeneXML entities.

In order to facilitate access to a database, a well-designed application programming interface is needed. A component of GeneX, the Perl module Genex.pm, allows each table of the database to be represented by a single Perl class in the GeneX namespace. Many of the classes represent GeneX-controlled vocabulary tables. There are also three utility modules to help with common data access and data conversion tasks, as well as Genex::DBUtils, a Perl utility for creating the SQL statements needed to interact with the database.

GeneX includes an HTML interface to the database that allows a user to run pre-set queries at several levels of complexity, as well as to select a data download format for further examination.

GeneX can be viewed as a set of analytical and visualization tools that eliminate much of the tedium of data assembly, access, and error checking. The prototype thus includes several statistical analysis methods and visualization tools that are described later. We thus hope to pique the interest of researchers who may contribute additional, and more novel, implements to the system.

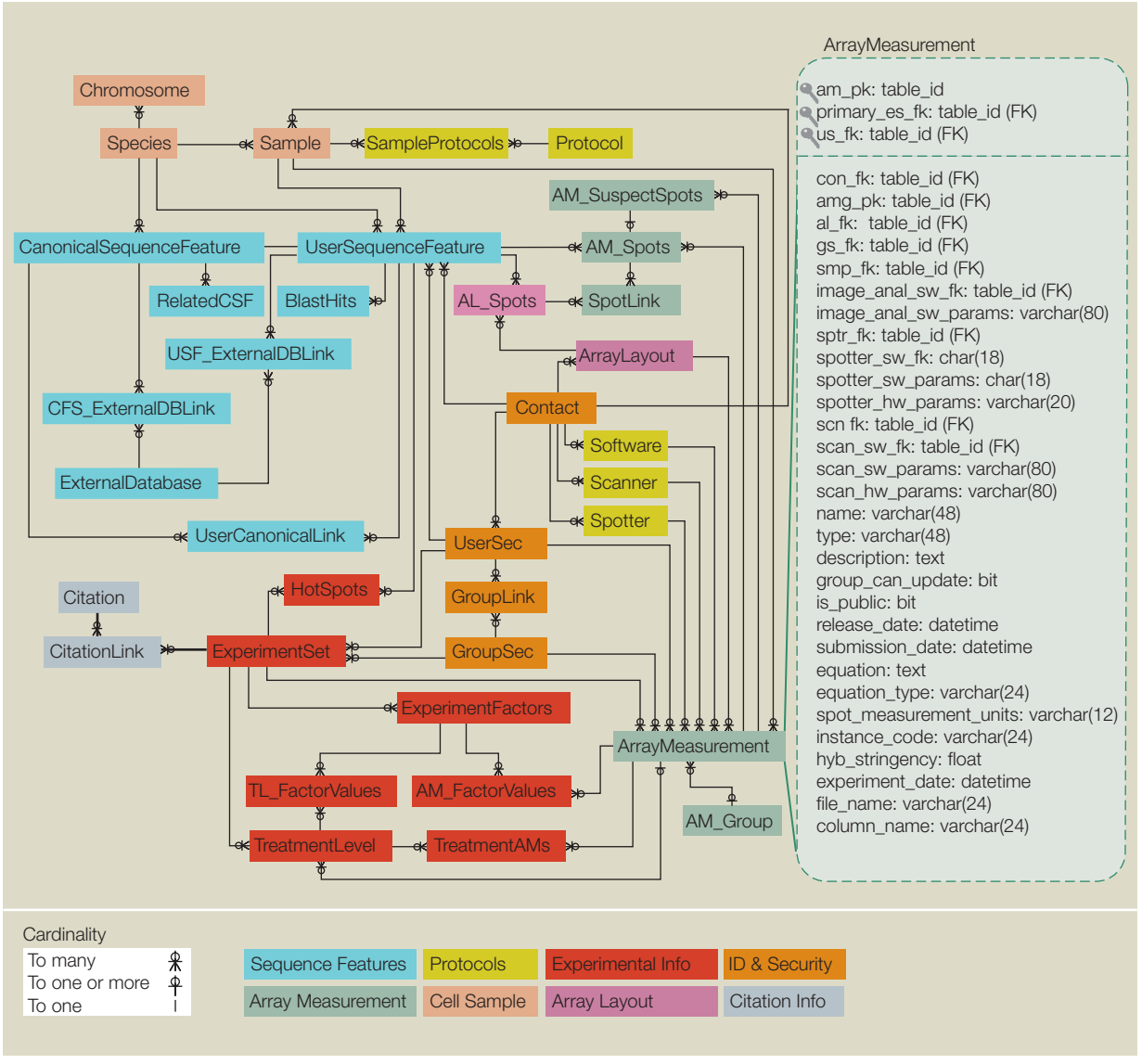
Because GeneX populates its internal sequence feature tables with relatively little sequence and sequence annotation data, the GeneX system supplies, where relevant, additional links to external databases so that users can query such domain-specific databases to obtain information. This makes detailed information in a given field available elsewhere accessible to GeneX users (although it cannot be used in complex queries since it is not integrated into the database). Once GeneX goes into wider use, we will make use of user feedback to guide us as to which of the sequence features are of most utility and should therefore be stored in the database (as opposed to storing links to external databases).

The Data Model

The Data Model must coherently and robustly represent the meta-data and their relationship to the primary experimental data, as well as the often-complex relationships among the meta-data objects themselves. In addition, it must represent the primary expression data at a sufficient level of abstraction to effectively make the data independent of the source technology. The GeneX Data Model is flexible enough to represent multiple signal channels, species, and technologies and we have made both the database schema and the implementation scripts available for examination, use, and critique.[29] This model provides a good starting point, which we hope will become more robust and scalable through the OSS peer review process.

Figure 2 illustrates the GeneX Data Model. Only a simplified diagram of the schema is shown here; for the full version see Reference 29. The functional roles of the various tables have been color-coded as indicated in the figure. The table names are largely self-descriptive with the following prefix abbreviations: CSF = Canonical Sequence Feature, USF =

Figure 2 The GeneX Data Model illustrated as an entity-relationship diagram



User Sequence Feature, AL = Array Layout, AM = Array Measurement, TL = Treatment Levels.

At the top of the hierarchy is ExperimentSet, each instance of which contains one or more ArrayMeasurements. Each ArrayMeasurement in turn contains one or more Spots. ArrayMeasurement can be seen as a description of a particular type of measurement from one or more Spots on an array, whereas the ArrayLayout provides the physical description of the array itself and therefore only one ArrayLayout is required for multiple ArrayMeasurements. Spots may have one or more values and therefore one or more ArrayMeasurements, depending on the parameters measured. For instance, a single Spot may have a measurement of its concentration, its background, or one of its experimental measurement acquisition channels. It may also represent a derived measurement such as a ratio. In such a case, there would be a separate row in the ArrayMeasurement table cor-

responding to each different measurement of the Spot. Logical groupings of ArrayMeasurements (e.g., Cy3 and Cy5 dye measurements from the same spot) are accomplished through the AM_Group table.

Meta-data. Experimental meta-data strongly influence the types of hypotheses generated. For example, the maize alcohol dehydrogenase1 (adh1) gene might be induced (the transcription rate of the gene is increased) by flooding in a flood-tolerant line of maize. In a second apparently similar experiment, in which a normal maize line was used, the same gene was apparently not induced at all. The hypotheses for explaining the observed difference might include only lineage (i.e., normal vs flood-tolerant) unless the investigator could also discover that, in the second experiment, a different technique for making the cDNA (complementary DNA) copy of the adh1 mRNA was used. If this technique failed to successfully produce the adh1 probe, the absence of signal has a methodological rather than a biological cause. GeneX captures extensive meta-data on many such levels. While complete descriptions of all levels are beyond the scope of this paper, two are detailed below for the purpose of illumination.

Associated with each ArrayMeasurement is one Sample element per channel of the technology used (e.g., Cy3/Cy5 cDNA microarrays[30] use two channels, whereas Affymetrix GeneChip** and Northern blot technologies[31] use a single channel). Each Sample contains detailed experimental information associated with the RNA source including the description of the source organism, biological treatment, and the protocols used for sample preparation and hybridization. Also associated with each ArrayMeasurement are a number of other types of meta-data, including descriptions of the hardware, software, and parameters used for generating, scanning, and analyzing the array.

Meta-data describing the "spot" (the DNA that has been immobilized on the substrate) consist of a detailed description of the spot's physical properties and the organism from which it was taken. GeneX distinguishes between information describing the actual physical DNA used (UserSequenceFeature—USF) and information describing that mRNA which the spotted DNA is intended to assay (CanonicalSequenceFeature—CSF). This allows GeneX to better reflect that in many cases, the material constituting the spot (a short oligonucleotide, for example) is quite different from the mRNA being assayed. The USF can also make use of fields within the Sample table to further describe source information such as developmental stage. The USF table links to the BlastHits table which stores the accession numbers and "expect" values of all similarity search matches the user deems significant. The CSF functions to limit ambiguities arising from non-uniform naming conventions. For each species, the CSF is populated with data that have been agreed upon by that research community. This allows the user to query for expression data based on specific gene names without having to have *a priori* knowledge of a particular laboratory's clone names, for example. Both the USF and CSF tables support links to external databases.

Storing analysis results. In addition to storing the numeric values for a spot measurement obtained from an instrument such as a scanner or optical densitometer (the "primary" or "raw" data), GeneX supports ratio and averaged data across ArrayMeasurements within an ExperimentSet. For some technologies (e.g., two-channel arrays, in which both the control and treatment probes hybridize to the same immobilized target DNA) the values are inherently controlled for each spot.

A ratio of the control to treated sample is typically produced as the fundamental value to be used in subsequent analyses in gene expression experiments. If replicates of a spot exist, the average of either the raw or the ratio values may be calculated and these averages become the fundamental value for subsequent analyses. Typically these ratios or averages reflect what the author of the data felt defined the "treatment." A treatment is here defined as those variables manipulated by the experimenter in the expectation that the biology of the test organism will be affected, leading to changes in transcription of certain mRNAs. Examples of treatments are heat stress, drought stress, varying amounts of nutritional supplements, etc. Treatment is distinguished from methodological and instrumental variables that are often outside the experimenter's control. These variables may affect the values measured, or the confidence that can be placed in them, but they are not expected to have affected the biology itself.

In many cases, there is more than one meaningful way to view the data, especially in multivariate experiments. The next iteration of the data model will support multiple views or "slices" through the ArrayMeasurements that make up an ExperimentSet and allow the creation of Virtual Experiment Sets from ArrayMeasurements belonging to different ExperimentSets. In addition, in the HotSpots table,

GeneX stores a list of all USFs within an ExperimentSet that exceed a threshold level of expression relative to control. This facilitates rapid retrieval of "interesting" data from any given ExperimentSet. This is based upon an assumption that the most interesting data are those for which a change of a certain magnitude is observed.

Comparison of data from different sources. Enabling meaningful comparisons of gene expression data across species is difficult, primarily due to a lack of a standard nomenclature and ontology. Without a common framework to define entities (such as genes) unambiguously, the results of such comparisons are of questionable value. One of the goals of the GeneX project is to stimulate discussion in the scientific community concerning the meaning of such comparisons, by developing cross-species controls and standardized gene comparisons. In addition to nomenclature issues, meaningful comparison of data from different laboratories presents difficulties that require discussion of the calibration and quantification standards necessary for every experiment. GeneX also faces the challenge of supporting the large variety of technologies used to generate expression data: Northern blot, cDNA microarray, Affymetrix' GeneChip,[32] Amplified Fragment Length Polymorphism[33] (AFLP), and SAGE. Although initially focusing on high-throughput, array-based technologies, the GeneX schema can support other technologies, although they are not currently supported by the Curation Tool. For example, Northern blot data are supported as a single ArrayMeasurement with a single spot. Further development of an accepted set of standards to calibrate results from different technologies is necessary.

Security. Security and data privacy are important attributes of any successful database. Access to data stored in the GeneX database can be defined as public or private on either an ExperimentSet or an ArrayMeasurement basis. Group access privileges allow owners of data to specify diverse access privileges to individuals on a per-ArrayMeasurement basis. Data are encrypted using the SSL during the network transfer, although other protocols could be used as well. Most of the communication, encrypted or not, uses standard port 80 on the Web server, because most organizations' firewalls are configured to pass all traffic on that port.

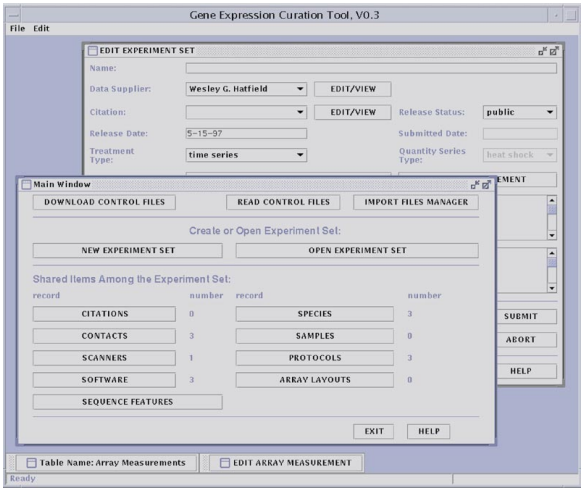
Implementation status. GeneX is designed for use either as a public resource located at NCGR, or as a locally installed database for individual labs to store and analyze their own data. NCGR is currently working with a number of microarray data producers to ensure that the GeneX data model remains flexible for both these applications. Because the GeneX database is still in its first year of operation and the amount of data accumulated is quite small, limitations in the scalability of the schema have not yet been determined. We assume that as the amount of data increases, certain bottlenecks will be found. To improve response time, we expect to enhance the hardware or streamline the schema, in which case, necessary scripts will be provided to users to migrate data to the new schema.

As an Open Source resource, users are free to change the locally installed schema to suit their purposes, but those changes will be merged into the NCGR-supported schema only after detailed review and discussion. Generally, additions will not affect the core schema, or the associated GeneXML, but those changes will not be supported in the XML or updates to the schema until they have been formally accepted. To merge additional data from user-modified schemas to refreshed NCGR schemas, the users will have to do some custom scripting. In order to support stability as well as encourage innovation, we may have to divide the development of GeneX, as is the practice in many OSS projects, and label versions with even and odd numbers. In this scheme, even numbers indicate stable versions, whereas odd numbers indicate unstable ones.

Curation Tool

Information about experimental conditions is often sparse and available only in hard copy—the lab notebook is still the most complete repository of experimental information in most laboratories. To help transfer the experimental data and associated metadata from the notebook to the database, and to assign much of the effort in doing so to the researcher rather than to a database curator, NCGR has developed a specialized Curation Tool.[34] Because the level of detail and user interaction required to organize and validate the data cannot be easily managed by simple HTML forms, the tool is written as a client-side Java application (see Figure 3). The primary goal of the Curation Tool is to ease the entry of new expression data, using restricted vocabularies and other curated data downloaded from the master GeneX site (note the buttons labeled DOWNLOAD CONTROL FILES and READ CONTROL FILES) to provide a point and click path through the process. The application behaves as a typical GUI (graphical user interface) ap-

Figure 3 Desktop view of the Curation Tool



plication and shields the user from manipulating GeneXML objects. Data are uploaded as a GeneXML item (Figure 1). The application also has the ability to read and write partial information and thus save data entry time by loading an existing record or even an entire experiment set as a template.

Controlled vocabulary lists are maintained by NCGR curators and requested automatically by the Curation Tool via a network connection upon activation. Very few of the Curation Tool's steps require user entry of new information; to maintain the controlled vocabulary, the user is encouraged to select one of the listed terms. In cases where necessary terms are not presented, the user may provide additional terms, which will be flagged for a curator to review upon receipt. Once the term is accepted, it is added to the controlled vocabulary list for future use. It is noteworthy that labs that have independent GeneX installations will be curating their own controlled vocabulary lists and can define their own vocabularies, which will make it easier for them to enter data that match their domain vocabulary. However, when they try to submit data to a higher-level database, the two vocabularies must be synchronized, which could lead to significant problems. To reduce this vocabulary collision, we are attempting to follow the framework proposed by the Gene Ontology Consortium,[35] whose members have investigated the vocabularies and ontologies of particular biological domains. We see user-supplied keywords as being useful as a sampling of the field, and these keywords can be forwarded to the Gene Ontology group for consideration.

The Curation Tool also allows researchers to archive collections of their original data with the submission so that when the data are needed again, they are available via a link from the database. This is provided as a convenience to users; these archives are stored on line or "near line," but external to the database, since they take up significant storage.

GeneX Markup Language

The GeneX system is designed to exchange data with other gene expression databases. A formal data exchange mechanism is especially important in gene expression analysis, because the experiments are expensive to perform and data analysis methods (especially normalization and standardization) are still in development. Exchanging complete data sets in standard format allows data to be reviewed and analyzed quickly. XML provides a standard for representing domain-specific structured data such as gene expression data. The GeneX team initiated the design of GeneXML[36] for this purpose, rather than a more compact, but also more obscure, binary format. NCGR is a member of a consortium, the MicroArray Gene Expression Database (MGED) group,[37] which is in the final stages of defining the Minimal Information for MicroArray Experiments (MIMAE) standard and the corresponding MicroArray Markup Language (MAML) to allow for more effective data exchange. The specifications will allow any supporting database to request and produce data in a format that another database or gene expression application can parse. The MGED group has submitted the MIMAE and MAML specifications to the Object Management Group's Life Sciences Task Force for Gene Expression for consideration to become a CORBA** (Common Object Request Broker Architecture**) standard. While NCGR's GeneXML is somewhat more technology-neutral than MAML, we will fully support MAML as an exchange format, once the specification is formalized, which was expected in March 2001 at the time this paper was written.

Data types associated with gene expression experiments. In addition to storing the numeric data from gene expression experiments, GeneX contains objects such as Contact, Citation, Scanner, Software, and Species, which capture the context of experiments. All meta-data are stored as header elements in a GeneXML document as specified by the docu-

ment type definition.[38] In addition, each element or label has a unique identifier (the XML equivalent of a pointer) that can be referred to by other elements (e.g., es_factor1 in GeneXML Listing 1[39]).

Gene expression data are most easily considered on a per-array basis, because this is the physical grouping of the data. Each array may have *N* channels of measurements. For example, the Cy3/Cy5 cDNA microarray[30] is a two-channel experiment. Multichannel technologies are under development by many organizations and we expect to see, in the near future, microarray variants with more than two channels of emission spectra. While the GeneX database schema addresses the multichannel data storage issue by grouping data as multiple array measurements, this method is not very intuitive in a context outside the database. In order to make the GeneXML structure more generic and therefore more useful for other database or analysis systems, GeneX groups the data on a per-array basis.

To support multichannel measurements, the *measurement_info* element in the array header lists and describes *N* numbers of measurements as well as a set of experimental treatment factors associated with each array (see GeneXML Listing 1[39]). In the data section, a spot on the array may then have as many measured values as correspond to the predefined measurements. Array data are organized by *feature*, where a feature is a specific piece of DNA sequence that has been fixed to the substrate. There may also be spot replicates of the same feature on one array (e.g., spot replicate controls). Under each GeneXML feature structure, there may be one or more spots with an identification reference to the array layout spot structure to indicate the distinct location of each spot replicate. Each spot then will have *N* measurements containing quantitative information such as raw intensity, background, corrected value, and ratio. This allows us to support not only the current Cy3/Cy5 microarrays, but also the common alternative technology platform using radioactive (1-channel) Southern[40] blot-like nylon arrays (see GeneXML Listing 2[41]).

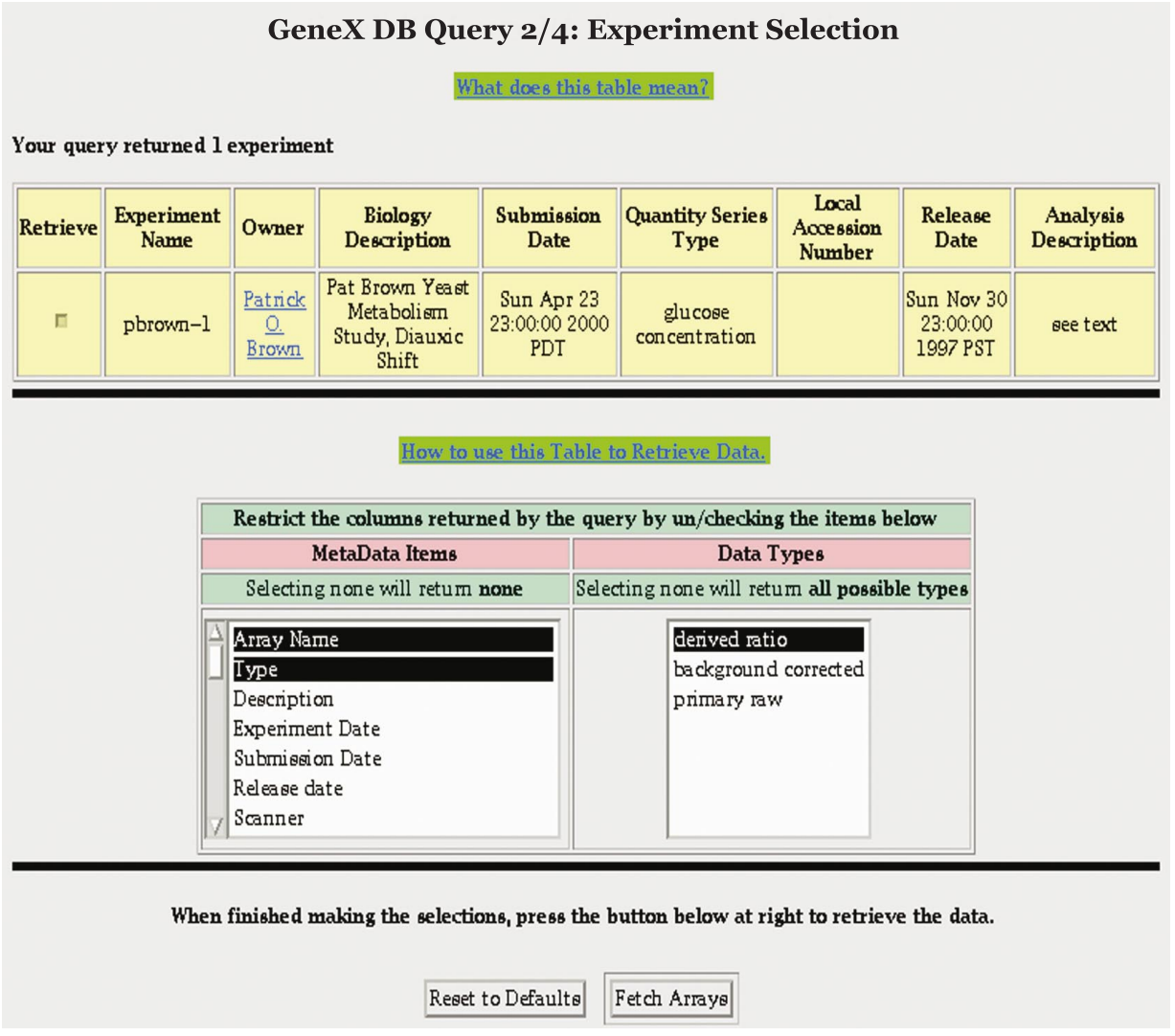
Numeric data handling. In order to structure numeric data, each number, representing a measurement of one spot per channel, is marked up with several bytes of information. The repetitive nature of this information introduces a size problem for a complex XML, with respect to both data download and parsing. In a test implementation formatted this way, a two-channel yeast experiment (approximately 6000 genes) with only seven arrays can require tens of megabytes (MB) of storage. Some Document Object Model XML parsers require the creation of in-memory hierarchies before the structure can be traversed, leading to prohibitive memory requirements. This problem will intensify in high-throughput situations in which hundreds of arrays comprise one experiment. The GeneX system remediates this problem by associating a separate, external expression values file for each array. The main GeneXML document now contains the meta-data, which require only tens of kilobytes. The sequence features that describe each piece of immobilized DNA used as a probe in the experiment and the array layout spots that describe the location of the spot information on the arrays are stored in two separate external files. For a 6000-feature array, each of the files is approximately one MB in size. The data files are stored separately for each physical array. For a two-channel yeast microarray experiment having 6000 features (with data for each including raw value, background, and corrected value), the per-array file size is about 2.5 MB. To maintain the coherence of a GeneXML data set, a list of these external files is specified in the main GeneXML document and is referenced by file identifiers.

Genex.pm

GeneX is distributed with the Perl module Genex.pm.[42] Genex.pm, which allows the user to manipulate the underlying database, makes use of the Perl DBI[43] (database interface) and the Perl XML::Parser[44] and XML::DOM[45] modules. The API to it is built with a "one class per table" approach, so that each table in the database is represented by a single Perl class in the Genex.pm namespace. A large subset of the classes represent the GeneX controlled vocabulary tables, each of which regulates the accepted values of a single column for a single table. Last, there are three utility modules for help with common data access and data conversion tasks. Genex::DBUtils is used for creating the SQL statements needed to interact with the database. To facilitate data input and output from GeneX in GeneXML format, we have included the Genex::XMLUtils module, as well as the Genex::HTMLUtils module to render Gene objects as HTML to display on Web pages.

Each module has accompanying documentation in both Perl *pod* (plain old documentation) format and HTML. Also supplied is a general overview of the API and a short tutorial for using the Genex.pm classes.

Figure 4 Phase 2 query screen



Each module also includes a set of automated regression tests, which allow testing of the entire system before installation and provide further details for using the Genex.pm API. Several examples[46] of code with annotations are available.

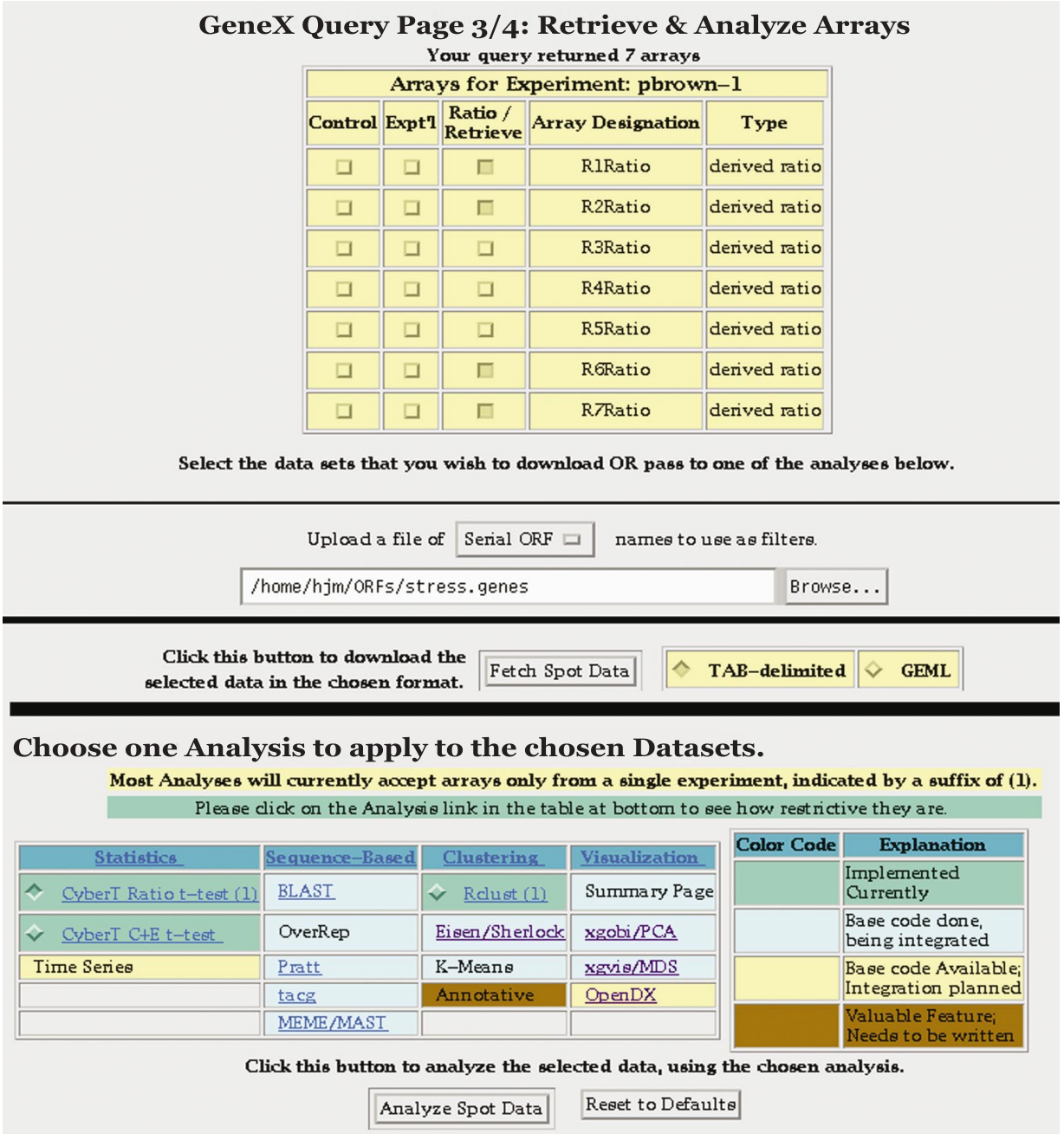
Query interface

GeneX includes an HTML interface to the database that allows users to:

1. Retrieve entire experiments as the result of broad queries (as shown in Figure 4)
2. Retrieve particular subsets of data as the result of more complex queries (e.g., a collection of data from selected timepoints across multiple experiments, as illustrated in Figure 5)
3. Be guided in making restricted queries returning consistent data sets appropriate for processing by a particular analysis routine
4. Download data in both tab-delimited and GeneXML format for storage on a local database or for analyses on the user's local machine

Examples are shown in Figures 4 and 5. Figure 4 shows the Phase 2 query screen stripped of browser

Figure 5 Phase 3 query screen



decorations. It contains a yellow HTML table representing an experiment summary built automatically by the Genex.pm HTMLUtils module, with a further filtering table below where the user can select several MetaData Items (lower left panel) to be retrieved along with the numeric Data Types (lower right panel). The HTMLUtils module also creates the hyperlinks in the experiment summary table to

point to further information that is implied by underlying foreign key relationships.

The Phase 3 query screen is shown in Figure 5, minus browser decorations and help links (for compactness) with the HTMLUtils-built table of arrays matching the query in Figure 4 displayed in yellow at the top. The arrays contain derived ratio data (the native data type for this experiment), which can be selected either as replicates as shown (sets 1, 2, 6, 7) or as representing Control and Experimental estimates. While this selection choice can easily be misused, it can also be used to make comparisons among arrays that would not otherwise be easily possible. Immediately below is a further filtering mechanism that can restrict the results based on a file of Gene or Open Reading Frame names (often, only a few of the thousands of genes that compose an entire data set are of interest). The user can select a download format in 2 tab-delimited formats or GeneXML, and (optionally) choose an analysis that can be applied to the selected data. The analyses are color coded to indicate implementation level.

HTML forms do supply a near-universal method of accessing the database, but are not very flexible for composing complex queries. Significantly more flexible query and visualization abilities can be made available using Java applets and applications. For example, we are communicating with the author of MaxdView to determine if it can be made compatible with GeneX.

Analytical routines and data visualization

The analysis of very large, internally complex data sets is a relatively new and active field in genomics. GeneX supplies several statistical routines and visualization tools for inspecting the results. As a separate but useful resource, one of the GeneX Web pages at the NCGR site includes a summary of gene expression databases, visualization tools, and analytical applications in a large Table of Gene Expression Resources.[47]

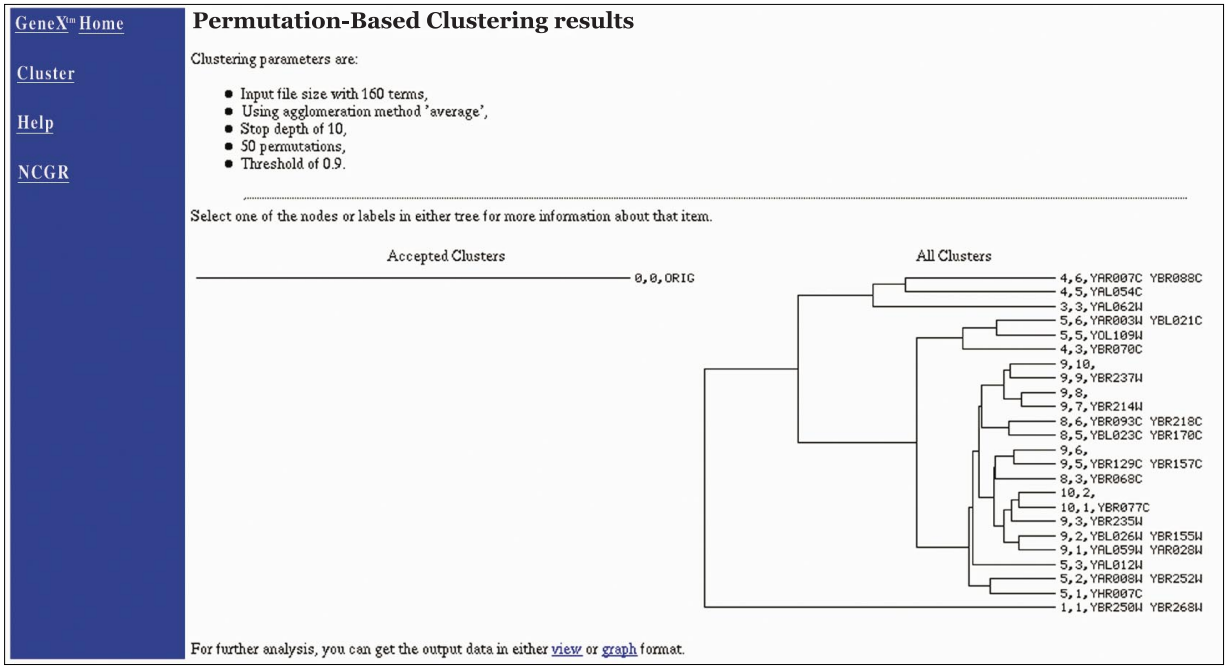
Clustering. Clustering routines are methods of organizing large sets of expression data into groups of genes sharing similar expression patterns. Because genes in the same cluster display coordinately regulated expression profiles, one can hypothesize that clusters represent genes with similar regulatory mechanisms. They may also represent genes that contribute to a similar functional response, for example, the detoxification of a pesticide, or change from the metabolism of one carbon source to another.[48]

Simple clustering procedures. The GeneX system includes a simple agglomerative hierarchical clustering procedure. The procedure employs two user options: a selection of six agglomeration methods and a stop-level parameter, which, when activated, terminates clustering at a user-defined level in the clustering tree (dendrogram). Hierarchical clustering routines provide grouping schemes at several levels: upper levels in the clustering tree represent coarser groupings of the data than those at lower levels. Halting the procedure at midlevel may thus alleviate uninformative separation of data elements. Also included is an interactive graphing tool that displays the clustering tree (see Figure 6).

Permutation-based clustering with associated threshold measures. Gene expression data contain considerable amounts of experimental error, which, upon application of clustering routines, could result in unreliable gene groupings. We believe that the association of a threshold measure to each cluster derived by a clustering procedure proves useful to the researcher. Permutation tests are used to simulate data sets with random gene expressions, providing a basis for comparison between clusters produced from experimentally obtained data and those resulting from randomly generated data sets. Based on this comparison, the threshold measure distinguishes between clusters that are statistically significant and those that are not, supplying an estimate of confidence in the clustering output.

The procedure assigns the threshold measure as follows. At each partition of the dendrogram, N permutations of the parent cluster are performed, generating N random data sets. Each random data set is partitioned by the same clustering routine, and a test statistic of each is computed. The test statistic of the original partition is then compared with the distribution of the N experimental statistics with respect to a user-specified threshold. Such a comparison provides an assessment of confidence in the inferred clusters. That is, if the measure of a parent cluster is lower than the user-defined threshold, its subclusters are not accepted (see Figure 6). The user may customize his or her clustering experiments by choosing one of several agglomerative clustering methods, declaring a specific number of permutations, and seeking an acceptance threshold. Although the user may employ as many permutations as desired, the procedure's run time increases significantly

Figure 6 The interactive graphing tool displays a clustering tree



as *N* increases. To compensate for this increased run time, the user may activate the stop-level parameter, thereby terminating the clustering process at a higher (and thus coarser) level.

The image in Figure 6 displays the original hierarchical clustering of a 160-gene data set (lower right), and its accepted clustering upon application of the permutation-based procedure with 50 permutations and a threshold of 90 percent (upper left). Note that none of the sub-clusters of the original data set is accepted, which indicates that the user can be quite confident that the original clustering is statistically insignificant. This differs considerably from most hierarchical clustering routines as it associates to each cluster a threshold measure, indicating how "sure" the researcher can be that the clustering is significant.

Externally developed clustering resources. We will also be adding other clustering approaches as they become available. For example, code based on Eisen's original hierarchical clustering routine[49] is now available for clustering across experiments as well as across genes, and Tamayo's Self-Organizing Map approach[50] is also included. K-means clustering is supported by both the Eisen-Sherlock xcluster and by Dysvik and Jonassen's J-Express,[51] which is bundled with GeneX. The latter also supports Principal Component Analysis, mosaic visualization (also known as Eisen maps), and simultaneous branch filtering of the cluster tree.

Visualization. Expression experiments often involve more than three variables so even static three-dimensional (3-D) plotting is inadequate. For graphics that can be reduced to 2-D or pseudo-3-D, GeneX uses the Open Source gnuplot.[52] For higher dimensional plotting, we are using xgobi[53] and xgvis,[54] two X Window System-based applications. We plan to add IBM's Data Explorer[55] in spring 2001. Data Explorer allows full three-dimensional manipulation of the data, permitting simultaneous display of thousands of points rendered in different colors, glyphs, and textures. These techniques are interactive, requiring both low latency and high bandwidth, and so are currently only available with local installations of GeneX.

Principal component analysis and multidimensional scaling. Many gene expression experiments are of high dimension (multiple treatments over time

courses, interactions, comparisons among cell types under different drug conditions). Therefore dimensional reduction, grand tours, correlation analysis, principal component analysis, and multidimensional scaling[56,57] will be of interest. These are currently supplied by xgobi, as well as by J-Express in a more visually compelling manner, but additional techniques (such as the related Support Vector Machine[58] approach) will also be added as users demand.

Integrated statistical system. R[22] is a near-clone of the S/SPlus[59] language, a popular modeling and analysis tool among statisticians. R supplies hundreds of internal and external routines for data analysis, including many types of clustering routines, and can export graphics directly to the user. Tony Long and Pierre Baldi (of the University of California, Irvine) have developed a set of R library functions useful for analyzing high-density array data that are already being used by bench scientists at the university to analyze high and medium density array data from *E. coli*, yeast, flies, and humans. This library, integrated into an application called CyberT,[60] performs repeated t-tests on the results from a control vs treated gene expression data set. It corrects for false positives by allowing users to set the Bonferonni correction value and can optionally perform a Bayesian estimation of variance[61] on the values in case of insufficient replicate data. It is also available in standalone form.[62]

Integration with other databases. GeneX has included links to many external databases and Internet resources. Some of these are species-specific (Saccharomyces Genome Database,[63] Mouse Genome Database,[64] The Arabidopsis Information Resource[65,66]), some more generic (dbEST,[67] GenBank,[68] Entrez,[69] KEGG,[70] EMBL,[71] ExPASy[72]). Species-specific databases are much richer sources of functional and developmental annotation than GenBank. The ability to search functional annotations would add to the subset of sequences whose expression data are considered appropriate to include for analysis. There have been advances in this area, driven by the need to automatically annotate the sequences from genome projects,[73] and in a few cases these annotation pipelines have been associated with gene expression projects as well as in Gaasterland's TANGO[74] system, now under development. GeneX currently supports this kind of annotative clustering weakly by supplying URLs (uniform resource locators) in context to other Web sites, but making the user intervene to mine the data available there; a tighter integration would be highly desirable.

Gene expression is a part of biological homeostasis, and for a gene expression database to be optimally useful, it should be integrated into a system that can address other important aspects of biology. We see GeneX as a significant part of a system that can integrate biological information from sequence to cell. This bioinformatic information pipeline starts with sequencing projects to capture the basic information available to a system, measures the expression of the genes in that genome, and finally relates the products of that expression to the metabolic processes that define the cell. While the GeneX project is still in its infancy, NCGR has two technologies that are applicable to this scenario. One is a sequencing pipeline based on the Phytophthora Genome Initiative[75,76] (PGI), which is a set of tools for capturing and tracking the information generated by a genome project, and the other is PathDB,[77] a database of metabolic pathway information that allows examination and simulation of formally defined pathways and allows pathway discovery as well. NCGR is currently starting to integrate these information engines with an integration technology developed in-house called ISYS**.[78,79]

Summary and conclusions

GeneX is an interactive database and analysis package prototype for storing, accessing, and analyzing gene expression data. It is designed not only to meet the needs of bench scientists but also of researchers who only need access to the experimental results. In either case the scientist may combine data produced by several laboratories. To ensure a general ability to do research, GeneX is designed to store and process nonspecies-specific data. This allows for data from many species to be analyzed together.

As new technologies for generating gene expression data are developed at a rapid pace, older technologies are quickly outdated. The GeneX Data Model is independent of technology type and thus the system is expected to sustain change well. The Data Model schema is implemented using popular relational database technology, so that as the relational technology improves, we expect to be able to transparently take advantage of it.

The GeneX Curation Tool assists the scientist in organizing, validating, and transferring data to a local or remote database. The Curation Tool is also a

mechanism to extract the perspective of the data generator as to the *de facto* ontology that the database is building.

The GeneX Markup Language is an XML type specific for exchanging gene expression data. Its purpose is to ease data communication with other gene expression databases. In particular, it fully supports MAML (MicroArray Markup Language).

The Genex.pm Perl module provides a high-level application programming interface for manipulating the database. Genex.pm also contains tools that use GeneXML as a transmission format for data import and retrieval from databases compatible with GeneX. The package also includes a tutorial and documentation.

A simple Web interface for querying the database has been written using the above-mentioned Genex.pm and HTML forms. A user may browse or download part or all of an experiment set. Additionally, blocks of data may be analyzed using one of several methods that are available on the GeneX server. Data are returned in a format suitable for the specific analysis routine.

GeneX provides a basic, Open Source infrastructure upon which others may build. The amount of gene expression data is growing exponentially and is expected to soon outpace the growth in sequence data. GeneX and the associated tool set are designed to help in organizing and storing these data. By providing a system that allows immediate inspection and analysis of the data, we hope to stimulate discussion about the best ways to perform the experiments and to analyze the data. By making the code publicly available, we hope other researchers will contribute code to GeneX and thus help us improve it.

Acknowledgments

We would like to thank our colleagues at NCGR for helpful comments and coding assistance during the development of GeneX, and in particular Peter Hraber, Andrew Tolopko, and David Bulmore. We are grateful for the feedback we have received from all our collaborators, and in particular the group at the University of California, Irvine: Suzanne Sandmeyer, Wes Hatfield, Tony Long, and Pierre Baldi. We thank the anonymous referees who reviewed our application to the National Science Foundation, and those who reviewed this paper for the *IBM Systems Journal*. We also wish to acknowledge the contribution of a talented contractor, Andrew Dalke, who helped code some of the Web scripts and provided insight into better security methods.

Cited references and notes

1. For a good introductory text on molecular biology, see B. Lewin, *Genes VII*, Oxford University Press, Oxford, UK (2000).
2. D. R. Bentley, "The Human Genome Project—An Overview," *Medicinal Research Reviews* **20**, No. 3, 189–196 (2000).
3. Except for the genomic rearrangement that accompanies the generation of immune diversity in higher multicellular organisms.
4. G. K. Geiss, R. E. Bumgarner, M. An, M. B. Agy, A. B. van't Wout, E. Hammersmark, V. Carter, D. Upchurch, J. I. Mullins, and M. G. Katze, "Large-Scale Monitoring of Host Cell Gene Expression During HIV-1 Infection Using cDNA Microarrays," *Virology* **266**, No. 1, 8–16 (2000).
5. D. T. Ross et al., "Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines," *Nature Genetics* **24**, No. 3, 227–235 (2000).
6. C. M. Perou et al., "Molecular Portraits of Human Breast Tumours," *Nature* **406**, No. 6797, 747–752 (2000).
7. U. Scherf et al., "A Gene Expression Database for the Molecular Pharmacology of Cancer," *Nature Genetics* **24**, No. 3, 236–244 (2000).
8. See http://www.rii.com.
9. See http://www.genelogic.com.
10. See http://www.ebi.ac.uk/arrayexpress/.
11. See http://bioinf.man.ac.uk/microarray/maxd/maxdSQL.
12. See http://www.ncbi.nlm.nih.gov/geo.
13. V. E. Velculescu et al., "Serial Analysis of Gene Expression," *Science* **270**, No. 5235, 484–487 (1995).
14. See http://genome.www4-stanford.edu/MicroArray/SMD.
15. See http://www.uk.research.att.com/vnc.
16. See http://www.gnu.org/copyleft/lesser.html.
17. See http://genex.sourceforge.net.
18. See http://www.w3.org/XML.
19. See http://www.apache.org.
20. L. Wall, R. L. Schwartz, and T. Christiansen, *Programming Perl*, 2nd Edition, O'Reilly & Associates Inc., Sebastopol, CA (1996).
21. See http://www.perl.com.
22. R. Ihaka et al., *The R Programming Language*, available from http://lib.stat.cmu.edu/R/CRAN/src/base.
23. K. Arnold and J. Gosling, *The Java Programming Language*, Addison-Wesley Publishing Co., Reading, MA (1996).
24. M. Stonebraker, L. A. Rowe, and M. Hirohama, "The Implementation of Postgres," *Transactions on Knowledge and Data Engineering* **2**, No. 1 (1990).
25. See http://www.postgres.org.
26. See http://www.ii.uib.no/~bjarted/jexpress.
27. See http://genome-www.stanford.edu/~sherlock/cluster.html.
28. See http://rana.lbl.gov/downloads/TreeView.zip.
29. See http://www.ncgr.org/research/genex/schema.shtml.
30. M. Schena et al., "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," *Science* **270**, No. 5235, 467–470 (1995).

31. J. C. Alwine, D. J. Kemp, and G. R. Stark, "Method for Detection of Specific RNAs in Agarose Gels by Transfer to Diazobenzyloxymethyl-Paper and Hybridization with DNA Probes," *Proceedings of the National Academy of Sciences (USA)* **74**, No. 12, 5350–5354 (1977).
32. R. J. Lipshutz et al., "Using Oligonucleotide Probe Arrays to Access Genetic Diversity," *Biotechniques* **19**, No. 3, 442–447 (1995).
33. J. W. Dekker and S. Easteal, "HLA-DP Typing by Amplified Fragment Length Polymorphisms (AFLPs)," *Immunogenetics* **32**, No. 1, 56–59 (1990).
34. The Curation Tool as well as supporting documentation and a tutorial for its use can be found at http://www.ncgr.org/research/genex/curation_tool.html.
35. See http://www.geneontology.org.
36. GeneXML was known as Gene Expression Markup Language (GEML) until it was discovered that a commercial entity (Rosetta Inpharmatics) also defined and trademarked their own GEML, a similar, although more limited, XML. In response, we have recently changed our specification name to GeneXML.
37. The group includes Stanford University, the Whitehead Institute, the University of Pennsylvania, the European Bioinformatics Institute, the German Cancer Research Center, the European Molecular Biology Laboratory, Incyte, GeneLogic, the University of Manchester, and others; see http://www.ebi.ac.uk/microarray/MGED.
38. See http://www.w3.org/XML/.
39. See http://www.ncgr.org/research/genex/IBMSJ.html#GeneXML_Listing_1.
40. E. M. Southern, "Detection of Specific Sequences Among DNA Fragments Separated by Gel Electrophoresis," *Journal of Molecular Biology* **98**, No. 3, 503–517 (1975).
41. See http://www.ncgr.org/research/genex/IBMSJ.html#GeneXML_Listing_2.
42. More information and source code available at http://www.ncgr.org/research/genex/genexml.html.
43. A. Descartes and T. Bunce, *Programming the Perl DBI*, O'Reilly & Associates, Inc., Cambridge, MA (2000).
44. The XML::Parser extension module is an interface to James Clark's XML parser, *Expat*; see http://www.w3org/TR/REC-DOM-LEVEL-1.
45. XML::DOM is a module written by Enno Derksen for building structures compliant with the Document Object Model (DOM) Level 1 specification; see http://www-4.ibm.com/software/developer/library/perl-xml-toolkit/.
46. See http://www.ncgr.org/research/genex/IBMSJ.html#Genex.pm_Examples.
47. See http://www.ncgr.org/research/genex/other_tools.html.
48. J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale," *Science* **278**, No. 5338, 680–686 (1997).
49. M. B. Eisen et al., "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Science of the USA* **95**, No. 25, 14863–14868 (1998).
50. P. Tamayo et al., "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation," *Proceedings of the National Academy of Sciences (USA)* **96**, No. 6, 2907–2912 (1999).
51. B. Dysvik and I. Jonassen, "J-Express: Exploring Gene Expression Data Using Java," *Bioinformatics*, in press.
52. T. Williams et al., *gnuplot*, available at http://www.cs.dartmouth.edu/gnuplot_info.html.
53. D. F. Swayne, D. Cook, and A. Buja, "XGobi: Interactive Dynamic Data Visualization in the X Window System," *Journal of Computational and Graphical Statistics* **7**, No. 1, 113–130 (1998).
54. A. Buja et al., "XGvis: Interactive Data Visualization with Multidimensional Scaling," *Journal of Computational and Graphical Statistics*, to be published in 2001.
55. Open DX.org, OpenDX, http://www.opendx.org/index2.php.
56. D. Asimov, "The Grand Tour: A Tool for Viewing Multidimensional Data," *SIAM Journal of Scientific and Statistical Computing* **6**, No. 1, 128–143 (1985).
57. A. Buja, D. Cook, and D. F. Swayne, "Interactive High-Dimensional Data Visualization," *Journal of Computational and Graphical Statistics* **5**, No. 1, 78–99 (1996).
58. C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery* **2**, No. 2, 121–167 (1998).
59. R. A. Becker, J. M. Chambers, and A. R. Wilks, *The New S Language*, Chapman & Hall, London (1988).
60. P. Baldi and A. D. Long, "A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-test and Statistical Inferences of Gene Changes," *Bioinformatics*, to be published in 2001.
61. P. Baldi, M. C. Vanier, and J. M. Bower, "On the Use of Bayesian Methods for Evaluating Compartmental Neural Models," *Journal of Computational Neuroscience* **5**, No. 3, 285–314 (1998).
62. See http://genomics.biochem.uci.edu/CyberT.
63. J. M. Cherry et al., "SGD: Saccharomyces Genome Database," *Nucleic Acids Research* **26**, No. 1, 73–79 (1998).
64. J. A. Blake et al., "The Mouse Genome Database (MGD): A Comprehensive Public Resource of Genetic, Phenotypic and Genomic Data, The Mouse Genome Informatics Group," *Nucleic Acids Research* **25**, No. 1, 85–91 (1997).
65. D. J. Flanders et al., "AtDB, the Arabidopsis Thaliana Database, and Graphical-Web-Display of Progress by the Arabidopsis Genome Initiative," *Nucleic Acids Research* **26**, No. 1, 80–84 (1998).
66. The Arabidopsis Information Resource is at the NCGR.
67. M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev, "dbEST—Database for 'Expressed Sequence Tags' " (letter), *Nature Genetics* **4**, No. 4, 332–333 (1993).
68. D. A. Benson et al., "GenBank," *Nucleic Acids Research* **28**, No. 1, 15–18 (2000).
69. T. A. Tatusova, I. Karsch-Mizrachi, and J. A. Ostell, "Complete Genomes in WWW Entrez: Data Representation and Analysis," *Bioinformatics* **15**, No. 7–8, 536–543 (1999).
70. M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research* **28**, No. 1, 27–30 (2000).
71. G. Stoesser, M. Moseley, R. Lopez, and P. Sterk, "The EMBL Nucleotide Sequence Database," *Nucleic Acids Research* **27**, No. 1, 18–24 (1999).
72. M. R. Wilkins et al., "Protein Identification and Analysis Tools in the ExPASy Server," *Methods in Molecular Biology* **112**, 531–552 (1999).
73. T. Gaasterland and C. W. Sensen, "MAGPIE: Automated Genome Interpretation," *Trends in Genetics* **12**, No. 2, 76–78 (1996).
74. See http://genomes.rockefeller.edu/research.shtml#tango.
75. See http://www.ncgr.org/research/pgi.
76. M. Waugh, P. Hraber, J. W. Weller, Y. Wu, G. Chen, J. Inman, D. Kiphart, and B. W. S. Sobral, "The Phytophthora Genome Initiative Database: Informatics and Analysis for Distributed Pathogenomic Research, *Nucleic Acids Research* **28,** No. 1, 87–90 (2000).
77. See http://www.ncgr.org/research/pathdb.

78. See http://www.ncgr.org/research/isys.
79. A. Siepel, A. Farmer, A. Tolopko, M. Zhuang, P. Mendes, W. Beavis, and B. Sobral, "ISYS: A Decentralized, Component-based Approach to the Integration of Heterogeneous Bioinformatics Resources," *Bioinformatics* **17,** No. 1, 83–94 (2001).

Accepted for publication December 14, 2000.

Harry Mangalam *National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505* (*electronic mail*: *hjm@ncgr.org*). Dr. Mangalam is a senior research scientist with the Gene Expression group. He received his doctorate from the University of California, San Diego, working with Geoff Rosenfeld on regulation of gene transcription, and he later found computer-aided sequence analysis strangely preferable to extended cold-room protein purification during a postdoctoral position at the Salk Institute. His research interests include distributed databases, data pipelines, sequence analysis, visualization, and societal aspects of Open Source software development.

Jason Stewart *OpenInformatics, 620 Arizona Street SE, Albuquerque, New Mexico 87108* (*electronic mail: jes@openinformatics.com*). Dr. Stewart is a bioinformatics consultant and programmer. He received his Ph.D. degree from the University of New Mexico in 1999 working with Ben Bederson on single display groupware (cooperative computer interfaces) using PAD++. His research interests include alternative human-computer interfaces, distributed bioinformatics databases, and deep Perl.

Jiaye Zhou *Inztro, 2521 San Pedro Drive NE, Suite F, Albuquerque, New Mexico 87110* (*electronic mail: jiaye@inztro.com*). Mr. Zhou is a computer science master student at the University of New Mexico, with an undergraduate background in biochemistry and molecular biology. He has been working with the GeneX group since December 1999. In addition to research in gene expression, his interests also include knowledge management systems, decision support systems, and distributed systems.

Karen Schlauch *Virginia Bioinformatics Institute, 1750 Kraft Drive, Suite 1100, Blacksburg, Virginia 24060* (*electronic mail: kas@vbi.vt.edu*). Dr. Schlauch is a senior research associate at the Virginia Bioinformatics Institute. She received her Ph.D. degree in 1998 from New Mexico State University in the field of mathematics. Currently, her efforts are mostly devoted to the development of novel analysis techniques of gene expression data. She is also involved in the development of mathematical models of gene networks.

Mark Waugh *National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505* (*electronic mail: mew@ncgr.org*). Mr. Waugh received his M.Sc. in plant molecular biology from New Mexico State University in 1991. He is a member of the broad-disciplinary science group at NCGR and has divided his time between the pathways and genomics programs. Most of his efforts have involved data model and interface design for several projects including metabolic pathways, gene expression, and sequencing pipelines. His current work involves automated sequence annotation with gene ontologies from the Gene Ontology Consortium.

Guanghong Chen *National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505* (*electronic mail: gc@ncgr.org*). Mr. Chen is a software developer who has worked in both the Gene Expression group and the Genome group. He received his master's degree in biology from the University of Notre Dame.

Andrew D. Farmer *National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505* (*electronic mail: adf@ncgr.org*). Mr. Farmer received a B.A. degree in philosophy and mathematics from St. John's College, Santa Fe, in 1993. He went on to begin a career in bioinformatics as a research assistant at the HIV database at Los Alamos National Laboratory. Following a position at the Santa Fe Institute where he applied hidden Markov models to genetic sequence alignments, he took his current position at NCGR, where he develops databases and software systems in bioinformatics.

Greg Colello *National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505* (*electronic mail: gdc@ncgr.org*). Dr. Colello, who received his doctorate in pharmacology from the University of Rochester in 1980, is the Technical Lead of the GeneX database project. His research interests also include XML for gene expression, plant and ecosystem simulation, and comparative genome mapping.

Jennifer W. Weller *Virginia Bioinformatics Institute, 1750 Kraft Drive, Suite 1400, Virginia Polytechnic Institute, Blacksburg, Virginia 24136* (*electronic mail: jwweller@vt.edu*). Dr. Weller is a research assistant professor at the Virginia Bioinformatics Institute at Virginia Tech and was formerly NCGR's program leader for gene expression. She earned a Ph.D. in biochemistry from the University of Montana, Missoula. Her current research includes development of an EST and genomic DNA analysis pipeline and a "wet lab" investigation of the genes expressed by *Arabidopsis thaliana* in root development and during parasitic attack.