© 2001 Nature Publishing Group http://genetics.nature.com

Minimum information about a microarray experiment (MIAME)—toward standards for microarray data

Alvis Brazma[1], Pascal Hingamp[2], John Quackenbush[3], Gavin Sherlock[4], Paul Spellman[5], Chris Stoeckert[6], John Aach[7], Wilhelm Ansorge[8], Catherine A. Ball[4], Helen C. Causton[9], Terry Gaasterland[10], Patrick Glenisson[11], Frank C.P. Holstege[12], Irene F. Kim[4], Victor Markowitz[13], John C. Matese[4], Helen Parkinson[1], Alan Robinson[1], Ugis Sarkans[1], Steffen Schulze-Kremer[14], Jason Stewart[15], Ronald Taylor[16], Jaak Vilo[1] & Martin Vingron[17]

Microarray analysis has become a widely used tool for the generation of gene expression data on a genomic scale. Although many significant results have been derived from microarray studies, one limitation has been the lack of standards for presenting and exchanging such data. Here we present a proposal, the Minimum Information About a Microarray Experiment (MIAME), that describes the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified. The ultimate goal of this work is to establish a standard for recording and reporting microarray-based gene expression data, which will in turn facilitate the establishment of databases and public repositories and enable the development of data analysis tools. With respect to MIAME, we concentrate on defining the content and structure of the necessary information rather than the technical format for capturing it.

© 2001 Nature Publishing Group http://genetics.nature.com

Introduction

After genome sequencing, DNA microarray analysis[1] has become the most widely used source of genome-scale data in the life sciences. Microarray expression studies are producing massive quantities of gene expression and other functional genomics data, which promise to provide key insights into gene function and interactions within and across metabolic pathways[2–4]. Unlike genome sequence data, however, which have standard formats for presentation and widely used tools and databases, much of the microarray data generated so far remain inaccessible to the broader research community.

Several factors contribute to the barrier to widespread access to microarray data. The field is young and has only recently approached the maturity needed to identify important aspects of the data. In addition, gene expression data are more complex than sequence data in that they are meaningful only in the context of a detailed description of the conditions under which they were generated, including the particular state of the living system under study and the perturbations to which it has been subjected. In contrast to an organism's genome, there are as many transcriptomes as there are cell types multiplied by environmental conditions. Moreover, comparing gene expression data is considerably more difficult, because at present, microarrays do not measure gene expression levels in any objective units. In fact, most measurements report only relative changes in gene expression, using a reference which is rarely standardized. Finally, different microarray platforms and experimental designs produce data in various formats and units and are normalized in different ways, all of which makes comparison and integration of these data an error-prone exercise[5,6].

Although the largest microarray laboratories have established their own databases[7], microarray data accompanying publications are typically reported on authors' web sites using a variety of formats, if they are accessible at all. Exactly what annotation should be provided for microarray data is open to debate, but it is clear that most of the publicly available data are currently not annotated in sufficient detail for use by independent parties (in fact, they are often not annotated at all). The reported data are often completely 'stripped' of all the evidence about the quality, reliability and possible error levels of particular data points. For instance, for two-channel microarray data, it is common to report only the background subtracted signal ratios without indicating anything about the absolute signal and background levels. Yet these are important for assessing the reliability of the measured expression for each arrayed gene.

[1]European Bioinformatics Institute, EMBL outstation, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [2]Centre d'Immunologie de Marseille Luminy (CIML TAGC) & Université de la Méditerranée, Marseille, France. [3]The Institute for Genomic Research (TIGR), Rockville, Maryland, USA. [4]Stanford University, Palo Alto, California, USA. [5]University of California Berkeley, Berkeley, California, USA. [6] University of Pennsylvania, Philadelphia, Pennsylvania, USA. [7]Department of Genetics, Harvard Medical School, Cambridge, Massachusetts, USA. [8]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [9]CSC/Imperial College School of Medicine Microarray Centre, London, UK. [10]Rockefeller University, New York, New York, USA. [11]Katholieke Universiteit Leuven, Leuven, Belgium. [12]University Medical Center, Utrecht, Netherlands. [13]GeneLogic Inc, Gaithersburg, Maryland, USA. [14]RZDP German Genome Resource Center, Berlin, Germany. [15]Open Informatics, Albuquerque, New Mexico, USA. [16]Center for Computational Pharmacology, University of Colorado School of Medicine, Denver, Colorado, USA. [17]Max Plank Institute for Molecular Genetics, Berlin, Germany. Correspondence should be addressed to A.B. (e-mail:brazma@ebi.ac.uk) or J.Q. (e-mail: johnq@tigr.org).

© 2001 Nature Publishing Group http://genetics.nature.com

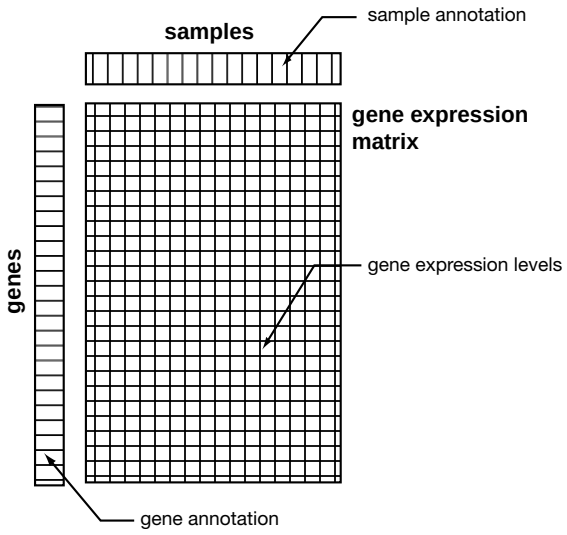

Fig. 1 Conceptual view of gene expression data. The model has three parts: (i) gene annotation, which may be given as links to gene sequence databases, (ii) sample annotation, for which there currently are no public external databases (except the species taxonomy) and (iii) the gene expression matrix, in which each position contains information characterizing the expression of a particular gene in a particular sample.

It is widely acknowledged that there is a need for public repositories for microarray data[8,9], whose functions would include providing access to supporting data for publications based on microarray experiments. Such repositories are under development by the National Center for Biotechnology Information (which has developed the Gene Expression Omnibus), the DNA Database of Japan, and the European Bioinformatics Institute (which has developed ArrayExpress); however, it is less clear exactly what information should be stored in such databases. Should databases store raw microarray scans (images), or is one final summary scalar per array element (such as one green/red ratio per spot for two channel platforms) sufficient? Or should some intermediate data, such as the complete output from a particular image analysis software package, be used instead? And should this be reported as 'raw' data or should it be normalized? What information about the experimental set-up should be required? And how should the array elements (printed spots or features) be annotated to facilitate an understanding of the experimental results?

The precise nature of the information to be stored will be dictated by the function of the particular database or repository. If the unique goal of the database is to archive supporting data for published experiments, it can be assumed that the publications themselves will provide information explaining the database entries. It may be argued that it can be left to peer review to ensure that a particular publication together with the respective database entry contains the information that is necessary to verify and reproduce the experimental results. It is unlikely, however, that such a system could be effective or scalable. Moreover, the value and usefulness of such a nonstandardized database would be considerably limited. For instance, it would be difficult to use the database for high-throughput automated data analysis or mining. The experience of the sequence databases over the past decade unequivocally demonstrates the strategic importance of structured, consistent annotation applied early in the process of data generation.

We believe that it is necessary to define the minimum information that must be reported, in order to ensure the interpretability of the experimental results generated using microarrays as well as their potential independent verification. Here we propose a document called MIAME, the Minimum Information About a Microarray Experiment, as a starting point for a broader community discussion. To make the task more manageable, we focus on microarray-based gene expression data, which arguably covers the most popular applications of microarray technology. We believe that the adoption of such a standard will facilitate the establishment and usefulness of microarray databases. MIAME should also prompt microarray manufacturers and software producers to develop adequate microarray laboratory information management systems (LIMS), enabling the production and capture of MIAME-compatible primary data at the bench.

The idea of having a defined minimum standard for information associated with experiments is not new to life sciences. A similar mode of operation has been adopted by the macromolecular structure community (see, for example, http://msd.ebi.ac.uk/), where most journals require submission of a well-defined minimum of raw data associated with publications. Not unlike crystallography data, those generated by microarray experiments are usually of a size and complexity that are meaningless to the general research community unless a minimum defined standard states what data are sufficient to support and verify conclusions.

Over and above representation of expression measurements, MIAME addresses the need for the comprehensive annotation necessary to interpret the results of microarray data. It is platform-independent but includes essential evidence about how the gene expression level measurements have been obtained. The first version of a MIAME document (MIAME 1.0) was recently completed and is used here as the basis for this discussion (Web Note A). A glossary of terms (Web Note B) and an example of a MIAME-compliant description of an experiment (Web Note C) were prepared to facilitate discussion of the proposed standard. MIAME is being developed by the Microarray Gene Expression Database group (MGED; http://www.mged.org/), a grass-roots movement to develop standards for microarray data[5]. MIAME 1.0 was approved in the MGED 3 meeting in Stanford University on 28–31 May 2001. It should be noted that MIAME does not specify the format in which the information should be provided, but only its content. A technical data format to capture microarray information in a form that includes MIAME requirements is being developed in a collaborative initiative coordinated by the Life Sciences Research Task Force of the Object Management Group (OMG) and, in fact, is well under way.

Gene expression—a conceptual view

A collection of gene expression data can be viewed abstractly as a table with rows representing genes, columns representing various samples and each position in the table describing the measurement for a particular gene in a particular sample (Fig. 1). We call this table a gene expression matrix. In addition to the matrix, a description of a microarray experiment should also contain information about the genes whose expression has been measured and the experimental conditions under which the samples were taken. The information required to describe a microarray experiment can be divided conceptually into three logical parts: gene annotation, sample annotation and a gene expression matrix (Fig. 1).

Ideally, we would like to measure amounts of gene expression in natural units, such as mRNA copies per cell[10], and to have an error estimate or reliability indicator such as the standard deviation (s.d.) associated with each value. There are a number of experimental challenges, however, that make direct measurement of gene expression difficult. Raw data from microarray experiments are images

© 2001 Nature Publishing Group http://genetics.nature.com

© 2001 Nature Publishing Group http://genetics.nature.com

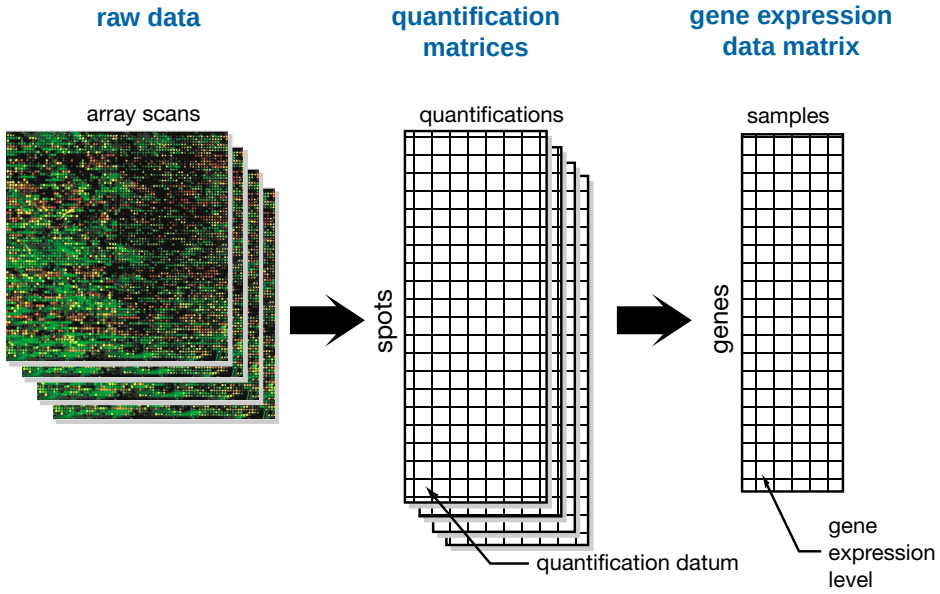

Fig. 2 Three levels of microarray gene expression data processing. The raw data from microarray experiments are images. These images have to be quantified by image analysis software, which identifies spots related to each element on the array and measures the fluorescence intensity of each spot in each channel, together with the background intensity and a number of other quantifications, depending on the particular software (microarray quantification matrices). To obtain the final gene expression matrix, all the quantities related to each gene (either on the same array or on replicate arrays) have to be combined and the entire matrix has to be normalized to make different arrays comparable.

© 2001 Nature Publishing Group http://genetics.nature.com

from hybridized microarray scans that have to be analyzed to identify and quantify each feature (spot) in the image. A DNA sequence may be spotted on a microarray several times; in addition, several distinct DNA sequences may be spotted that map to the same gene. To yield a single value for these, the corresponding measurements have to be combined. Moreover, the same biological condition can be measured in several (replicate) hybridizations, and the information from all replicates has to be summarized to derive a single gene expression data matrix. Finally, to compare microarray data from several samples, the data must be appropriately normalized.

There are at least three levels of data relevant to a microarray experiment: (i) the scanned images (raw data); (ii) the quantitative outputs from the image analysis procedure (microarray quantitation matrices); and (iii) the derived measurements (gene expression data matrices; Fig. 2). There is an important series of transformations leading from raw data to the gene expression matrix, and the steps involved are far from being standardized.

As there are no widely used standard controls for microarray assays, microarray data from different sources use different measurement units whose conversion factors are typically unknown and may even vary depending on expression level. This indicates the necessity to record not only the final gene expression matrix, but also a detailed description of how the expression values were obtained, if verification of the data is to be ensured. Consequently, the nature of the data that must be recorded necessarily becomes more complex.

Because microarray data have meaning only in the context of the particular biological sample and the exact conditions under which the samples were taken, a major element of the standard that we propose addresses sample annotation. For instance, if we are interested in finding out how different cell types react to treatments with various chemical compounds, we must record unambiguous information about the cell types and compounds used in the experiments. This information should be contained in sample annotation.

Although gene annotation can to a certain extent be expressed by links to sequence databases, the possibly complicated many-to-many relationships between genes in the gene expression matrix and elements on the array make it necessary to provide a full and detailed description of each element on the array.

General principles of MIAME design

As a starting point, we propose that for the data and annotations from microarray experiments to have the most value, they should satisfy the following requirements: (i) the recorded information about each experiment should be sufficient to interpret the experiment and should be detailed enough to enable comparisons to similar experiments and permit replication of experiments and (ii) the information should be structured in a way that enables useful querying as well as automated data analysis and mining.

The first requirement implies that a detailed annotation of the sample and other experimental conditions should be recorded and that some reliability estimates of particular data points should be given. For example, red/green ratios alone for two-channel platforms cannot normally be regarded as sufficient—currently there is no widely accepted method for indicating the confidence in a measurement, and much intensity-specific and expression-level information is lost. The necessary level of detail and whether the raw image data should be included are less obvious and are still widely discussed in the microarray community.

The second requirement implies the need for controlled vocabularies and ontologies to represent data as well as the need to limit free-format text only to cases where more structured representations are not feasible. This includes the use of a standardized nomenclature for the description of biological samples and conditions. The nomenclature may be as simple as a controlled vocabulary or as fully developed as an ontology. Usage of the same taxonomic classification (http://www.ncbi.nlm.nih.gov/Taxonomy/) is an example of an ontology that allows researchers to unambiguously determine the phylogenetic relationship between organisms. Unfortunately, such resources are not available for many other types of sample annotation, particularly those specific for individual species, such as 'developmental stage' (see the MGED Ontology Working Group home page for reference). A practical approach may be to initially use free-text descriptions for some sample annotations, despite the difficulty of incorporating them in automated queries. The use of free text may be sufficient, for example, in describing details of a laboratory protocol. A reference to a publication describing the experiment is an alternative; however, there are obvious drawbacks to not having information in hand for either queries or browsing.

Although the goal of MIAME is to specify only the content of the information and not the technical format, MIAME includes recommendations for which parts of the information should be provided as controlled vocabularies. The distinction between free-text format and controlled vocabularies influences the

© 2001 Nature Publishing Group http://genetics.nature.com

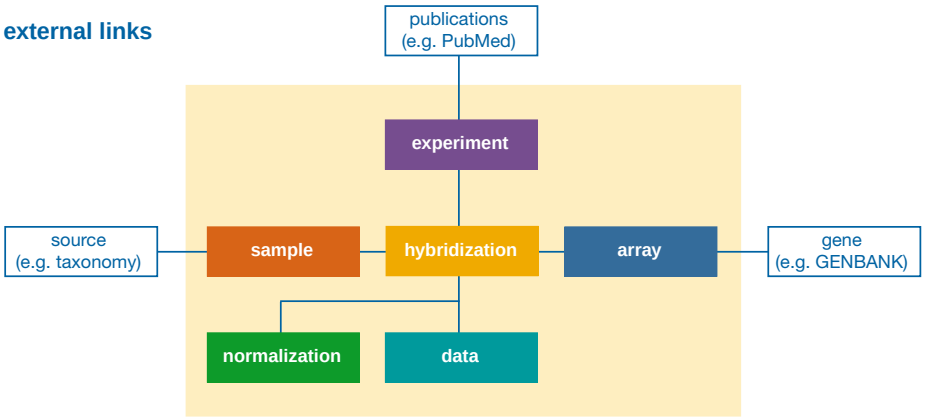

Fig. 3 A schematic representation of six components of a microarray experiment.

information content: a defined term taken from a given controlled vocabulary is more precise in its meaning than the same term used in a free-text field and can provide more advanced data query and analysis options. As the majority of the necessary controlled vocabularies do not exist, the MIAME definition includes lists of 'qualifier, value, source' triplets, which authors can use to define their own qualifiers and provide the appropriate values. For instance:

qualifier: cell type
value: epithelial
source: Gray's anatomy (38th ed.)
or
qualifier: treatment
value: 15′ heat shock
source: Smith and Jones, *Nature Genet.* (1992)

Given sufficient detail by the author, these triplets can fully describe a particular aspect of an experiment. The idea stems from the information sciences, where a 'qualifier' defines a concept and a 'value' contains the appropriate instance of the concept. 'Source' is either user-defined or a reference to an externally defined ontology or controlled vocabulary, such as the species taxonomy database. The judgment regarding the necessary level of detail is left to the data providers. In the future, these qualifier lists may be gradually supplemented with predefined fields as the respective ontologies are developed.

The aim of establishing microarray databases should be kept in mind, although the MIAME document is conceptually independent from it. An important principle in MIAME is that its parts can be provided as references or links to pre-existing and identifiable descriptions. For instance, for commercial or other standard arrays, required information needs to be provided only once by the array supplier and referenced thereafter by the users. Standard protocols also need to be provided only once after they are established, whereas specific deviations and parameters may be provided with each experiment (note that this would allow users to create a library of standard protocols). It is necessary that either a valid reference or the information itself be provided for every experimental data set.

There is one important additional principle underpinning MIAME: as microarray technology is developing rapidly, it would be counterproductive and unrealistic to impose on users any particular platform, software or methods of data analysis. Instead the standards should simply require the description of data in sufficient detail and with sufficient annotation, so that interested parties will have all the necessary information to understand how conclusions were reached. Note that we assume that data will be produced by different experimental platforms and laboratories, which means that few default assumptions can be made and most of the information should be reported explicitly by each laboratory.

In developing MIAME, we have sought to find a compromise between placing a burden on data producers to annotate experiments in elaborate detail and ensuring that data are annotated in enough detail to be useful to the general research community. Too much detail may be too taxing for data producers and may complicate data recording and database submission, whereas too little detail may limit the usefulness of the data. MIAME is an informal specification, the goal of which is to guide cooperative data providers. It is not designed to close all possible loopholes in data submission requirements. MIAME is not designed as a 'questionnaire' that can be filled in, but only as an informal specification on which microarray experiment–annotation tools may be based.

The six parts of MIAME

We define a microarray experiment as a set of one or more hybridizations, each of which relates one or more samples to one or more arrays. The hybridized array is then scanned and the resulting image analyzed, relating each element on the array with a set of measurements (Fig. 3). The data are normalized and combined with data from replicate hybridizations.

The minimum information about a published microarray-based gene expression experiment includes a description of the following six sections:

1. Experimental design: the set of hybridization experiments as a whole
2. Array design: each array used and each element (spot, feature) on the array
3. Samples: samples used, extract preparation and labeling
4. Hybridizations: procedures and parameters
5. Measurements: images, quantification and specifications
6. Normalization controls: types, values and specifications

Each of these sections contains information that can be provided using controlled vocabularies, as well as fields that use free-text format. Here we discuss only the general information required in each of these sections; for a full description, see the MIAME document (Web Note A), which includes a sample experiment described according to MIAME requirements. We do not discuss why we regard each of the MIAME elements as necessary, but we hope that this follows from the principles discussed in the previous sections.

In constructing the MIAME standard, we were careful to include as much relevant information as possible to aid in the interpretation of the results of each microarray experiment. Some have suggested that this is excessive as, for example, sample preparation and labeling protocols typically appear in the Methods sections of publications associated with microarray experiments. Although these

© 2001 Nature Publishing Group http://genetics.nature.com

© 2001 Nature Publishing Group http://genetics.nature.com

© 2001 Nature Publishing Group http://genetics.nature.com

assertions are merited, we believe that the comprehensive nature of MIAME confers distinct advantages without imposing an excessive burden on submitters of microarray data. There are a number of reasons for this. First, stand-alone entries make the use of a database inherently more efficient. Second, a collection of protocols in a relatively standard format (including controlled vocabularies) would facilitate comparison and usage of the protocols by third parties. Third, not all journals publish experimental protocols in sufficient detail (often due to length considerations) for others to reproduce them. Fourth, once this information has been prepared for a journal publication, providing it for a database submission should not constitute significant additional effort. Moreover, journals are increasingly relying on electronic release of both protocols and the data supporting published reports, and the MIAME standard would allow for a uniform presentation of such information.

Part 1: Experimental design. This section describes the experiment as a whole, which may consist of one or more hybridizations. Normally an 'experiment' should include a set of hybridizations that are inter-related and address a common biological question, such as all hybridizations relating to research published in a single paper. Each experiment should have an author (submitter) as well as contact information, links (URL), citations and a single-sentence experiment title. The section also includes a free-text format description of the experiment or a link to an electronically available publication.

The minimal information required in this section includes the type of the experiment (such as normal-versus-diseased comparison, time course, dose response, and so on) and the experimental variables, including parameters or conditions tested (such as time, dose, genetic variation or response to a treatment or compound). This section also provides general quality-related indicators such as usage and types of replicates and quality-control steps (such as dealing with low-complexity sequence-induced nonspecific hybridization). Provided in a format of controlled vocabularies, these will enable accurate queries and more formal data analysis than free-text descriptions.

Finally, this section specifies the experimental relationships between the array and sample entities—that is, which samples and which arrays were used in each hybridization assay. Each of these will be assigned unique identifiers that are cross-referenced with the information provided in the following sections. This information will allow the user to reconstruct unambiguously the experimental design and to relate together information from further MIAME sections.

Part 2: Array design. The aim of this section is to provide a systematic definition of all arrays used in the experiment, including the genes represented and their physical layout on the array. There are two parts to this section. The first is a list of the physical arrays; each member of the list is a simple description that gives a unique ID to each array used in the experiment and a reference to a particular array design. These designs are described in the second part of the section.

In the context of a database, array types should be defined and submitted only once by the array provider and referred to thereafter by users of the arrays. The array-type definition includes information common to all arrays of a particular type (such as glass-slide spotted with PCR-amplified cDNA clones) as well as precise descriptions of the physical content of each element (spot or feature). This section consists of three parts: (i) a description of the array as a whole (such as platform type, provider and surface type); (ii) a description of each type of element or spot used (properties that are typically common to many elements, such as 'synthesized oligo-nucleotides' or 'PCR products from cDNA clones'); and (iii) a description of the specific properties of each element, such as the DNA sequence and, possibly, quality-control indicators.

The challenge for element definition is to achieve a unique and unambiguous description of the element. Because references to an external gene index may not be stable, it is essential to physically identify each element's composition. Disclosing the nature of the relationship between an array element and its cognate gene's transcript allows informed assessment of an element's potential for nonspecific cross-hybridization or its capacity to distinguish alternative splice variants. Thus, where elements are based on cDNA clones, PCR amplicons or composite oligonucleotides, it is necessary that clone IDs, primer pair sequences or oligonucleotide sequence sets, respectively, are specified. In the case of commercial arrays where such details may be proprietary, MIAME allows for a compromise: instead of the actual probe sequence, the reference sequence from which the probe was derived may be specified. Although we do not consider this to be ideal, as the probe sequence may affect the results and interpretation of hybridization, it does allow commercial organizations to protect their intellectual investment in developing their array reagents while providing the minimum information necessary to uniquely identify the array elements.

Part 3: Samples. This section describes the second partner in the hybridization reaction: the labeled nucleic acids that represent the transcripts in the sample. The MIAME 'sample' concept represents the biological material (or biomaterial) for which the gene expression profile is being established. This section is divided into three parts which describe the source of the original sample (such as organism taxonomy and cell type) and any biological *in vivo* or *in vitro* treatments applied, the technical extraction of the nucleic acids, and their subsequent labeling.

As the characteristics necessary to accurately define a biological sample vary greatly from organism to organism, most of the biological sample definition is provided as an adaptable list of 'qualifier, value, source' triplets (such as 'strain', '129P1-Lama2^{dy}' or 'ICSGNM'). Currently, the single common feature of all samples is the organism's taxonomic definition. A list of qualifiers initially left at a submitter's discretion may progressively be made standard when applicable ontologies are made public.

As for laboratory protocols for sample treatments, sample extraction and labeling, these will need to be specified initially as free-format text. Again, it is anticipated that popular protocols will be provided once and referred to thereafter by submitters pointing out the exact parameters and deviations from the standard protocol. Knowledge of these protocols may be important for interpreting the data.

Part 4: Hybridizations. This section defines the laboratory conditions under which the hybridizations were carried out. Other than a free-text description of the hybridization protocol, MIAME requires that a number of critical hybridization parameters are explicitly specified: choice of hybridization solution (such as salt and detergent concentrations), nature of the blocking agent, wash procedure, quantity of labeled target used, hybridization time, volume, temperature and descriptions of the hybridization instruments.

Part 5: Measurements. The actual experimental results are defined in this section. It consists of the three parts discussed in Section 2, progressing from raw to processed data: (a) the original scans of the array (images), (b) the microarray quantification matrices based on image analysis, and (c) the final gene expression matrix after normalization and consolidation from possible replicates.

Image data should be provided as raw scanner image files (such as TIFF), accompanied by scanning information that includes relevant scan parameters and laboratory protocols. MIAME does not require a particular image format, only that submitters provide the original scans upon which data quantification was based

© 2001 Nature Publishing Group http://genetics.nature.com

© 2001 Nature Publishing Group http://genetics.nature.com

in a format readable by generally available software. Storing the primary image files would require a significant quantity of disk space, and there is no community consensus as to whether this would be cost-effective or whether this should be the task of public repositories or the primary authors. Nevertheless, as images represent the primary data from a microarray assay and the algorithms used for analysis can affect the conclusions that are reached, the current MIAME standard includes a specification for image deposition. As scanning protocols and image analysis methods mature, this mandatory requirement on image files may be revisited.

For each experimental image, a microarray quantification matrix contains the complete image analysis output as directly generated by the image analysis software (normally provided as separate spreadsheet-type files). Note that for a given image this is a 2D matrix, where array elements (spots or features) constitute one dimension and quantification types (such as mean and median intensity, mean or median background intensity) are the second dimension. We also provide in this section the co-lateral information needed to understand how image analysis was carried out, in particular the software used, the underlying methodology (such as algorithms and statistics), all relevant parameters and the definitions of the quantifications used (such as mean or median intensity). Note that if authors use their own custom-made (or customized) image-analysis software, the specification of its output is not formally dictated by MIAME. Nevertheless, in the spirit of MIAME, the output should include the information that permits the nature and quality of individual spot measurements to be assessed.

Finally, the gene expression matrix (summarized information) consists of sets of gene expression levels for each sample. If microarray quantification matrices can be considered spot/image centric, then the gene expression matrix is gene/sample centric. At this point, the expression values may have been normalized, consolidated and transformed in any number of ways by the submitter in order to present the data in a form amenable to scientific analysis. Rather than attempting to impose a standard for gene expression values, MIAME indicates preferred detailed specifications of all numerical calculations applied to unprocessed quantifications in (b) that have led to the data in (c). Experimenters are encouraged, though not required, to provide reliability indicators (such as s.d.) for each data point.

Part 6: Normalization controls. A typical microarray experiment involves a number of hybridization assays in which the data from multiple samples are analyzed to identify relative changes in expression levels, identify differentially expressed genes and, in many cases, discover classes of genes or samples having similar patterns of expression. A typical experiment follows a 'reference design' (more sophisticated loop designs have been proposed[11], although these have not yet been widely adopted) in which many samples are compared to a common reference sample so as to facilitate inferences about relative expression changes between samples. For these comparisons, the reported hybridization intensities derived from image processing must first be normalized. Normalization adjusts for a number of technical variations between and within single hybridizations, namely quantity of starting RNA and labeling and detection efficiencies for each sample. There are a variety of normalization schemes in use, including total-intensity, ratio-based and both linear and nonlinear regression techniques. In addition, these analyses may be based on either the complete data set, a user-defined subset of genes (often a set of 'housekeeping genes' thought not to change their level of expression under the conditions used) or exogenous genes for which RNA is 'spiked' into the initial samples of interest. Whether used for normalization or not, the use of exogenous controls is becoming increasingly common both for quality control within single arrays and for array-hybridization comparisons within and between platforms.

Section 6 of the MIAME standard provides an opportunity for the specification of parameters relevant to normalization and control elements. Our proposed standard includes (i) the normalization strategy (spiking, housekeeping genes, total array, other approach) (ii) the normalization and quality control algorithms used, (iii) the identities and location of the array elements serving as controls, as well as their type (spiking, normalization, negative or positive hybridization controls, 'landing lights' to assist spotfinding), and (iv) hybridization extract preparation, detailing how the control samples are included in sample targets prior to hybridization.

Discussion

Our goal is to develop a standard that can serve both research scientists and software developers. To that end, we hope that this description will stimulate discussion of the proposed MIAME standards and we encourage the microarray community, as well as the general research community, to provide us with their views on how this standard can be improved. For this purpose an e-mail discussion group has been set up by MGED consortium (to join, see http://www.mged.org).

At first glance, the extent of the information requested in the MIAME specification may seem daunting. It should be noted, however, that for most laboratories the majority of the information will be similar for many experiments, and once that information is specified, it should not have to be specified again. For example, most laboratories will use a single design for tens or hundreds of microarray experiments. The same is true of labeling protocols and normalization strategies. Describing these and other specifications has several goals: to help scientists conducting and designing experiments to record appropriate data, to assist those scientists interpreting or analyzing those data and to facilitate the design of databases and software that enable the data to be archived, queried and retrieved in an intuitive and biologically relevant manner.

We pose several questions to the research community. Is the current MIAME draft sufficiently detailed to capture the information needed to analyze and evaluate microarray data? If not, what is missing? Or is MIAME already too extensive, requiring specific details that are unlikely to ever be exploited; if so, what is superfluous? Are there objections to having a defined minimum information standard in principle and, if so, what are the alternatives? The goal of our proposal is not to impose specific solutions upon the community but instead to establish a community-wide understanding of the optimal infrastructure for the sharing of microarray data. It is possible that parts of MIAME may be burdensome whereas other sections do not offer sufficient detail. One important consideration is the queries that one would like to make of a MIAME-supportive gene expression database. Our development of MIAME was guided primarily by a desire to provide the information necessary to make such queries.

The MIAME document represents an overall consensus of the MGED working group on microarray data annotations[1] in all parts except section 5(a) concerning 'hybridization-scan raw data'. A majority of the working group supports the view that providing raw image data is an essential part of MIAME. There is also a considerable minority, however, who do not adhere to this view. If the consensus emerges that the primary image data are important, what is the preferred mechanism for ensuring access to images? Should they be stored in public repositories, or should the availability of the images be the responsibility of the experimenter? We anticipate that the answer to this (and other questions posed here) may evolve over time.

© 2001 Nature Publishing Group http://genetics.nature.com

© 2001 Nature Publishing Group http://genetics.nature.com

A more fundamental question for discussion is whether natural units for gene expression measurements exist. If so, what might they be, and can they be calculated from microarray measurements? In the absence of natural units of gene expression, how should gene expression data be organized, in particular to facilitate cross-experiment and cross-platform transcriptome comparisons? Is it possible and helpful to introduce standard controls and protocols for microarray experiments themselves, to facilitate comparison of the data?

Once MIAME has stabilized and a general consensus is reached, we can turn to practical applications. An initial technical application is to develop a data model that is able to record MIAME. Such a data model is already being developed within the OMG with the participation of the MGED consortium (http://www.geml.org/omg.htm). Founded on this data model is a standard data-exchange format (an XML description called MAGE-ML: Microarray Gene Expression Markup Language), which will allow communication of MIAME supportive data between local laboratory databases, central archives and standalone analysis packages. The final version of the MAGE-ML standard has been submitted to OMG, and participating organizations are already concentrating on the development of the supporting software.

The next important step is the development of gene expression databases able to record MIAME information and, equally important, of data submission tools. Such tools may include web-based questionnaires allowing users to enter MIAME information directly into a database or to export captured data in the standard format discussed above. In many cases, it is envisaged that most of the MIAME information will be recorded through local LIMS software before being uploaded into central archiving databases using a standard data-exchange format. As such, the development of such MIAME-friendly LIMS software will be an important task. We hope that by adopting and publicizing MIAME, we will encourage software developers to adapt their tools to this standard.

It is important that the effort to define minimum information requirements and data-exchange formats is endorsed by many major commercial genomics and bioinformatics companies. The availability of minimum data requirements will help in developing databases that can exchange information with public or other private databases. MIAME addresses the problem faced by most commercial gene expression companies of integrating gene expression data from multiple sources and multiple platforms.

Eventually, when MIAME–supportive public repositories are established, the general research community must consider whether full data disclosure should be required for publication. Journals and funding agencies will also have to consider whether, in the tradition of DNA sequence and macromolecular structure data, release of microarray data at a MIAME compliant level should be required.

In its present form, MIAME 1.0 is the first version of a document describing the minimal information required to report an array based gene expression experiment. Although some of the current specifications may become redundant or irrelevant as the technology evolves, extra information may need to be added at a later date. We therefore plan to couple future versions with progress in technology and analysis as well as experience gained within the microarray community. In addition, microarrays can be used for many types of experiment other than monitoring gene expression (comparative genome hybridization, genome mismatch scanning, chromatin IP experiments, and so on), and future versions of MIAME will attempt to accommodate these other types of data.

During this era of genomic-scale experiments, establishing expectations for format and content, sharing data and analysis tools and establishing databases and other resources has become a widespread problem throughout the life sciences. For instance, the neuroimaging community seems to be confronted with very similar problems (how to compare data across different laboratories, a lack of standards for data normalization, a need for standard annotations) and is following a similar strategy for developing a solution[12]. Our hope is that such an approach becomes the norm by which data presentation and publication standards are developed in the future. As such, we look forward to hearing comments and suggestions from the general research community.

Note: Supplementary information is available on the Nature Genetics web site (http://genetics.nature.com/supplementary_info/).

Acknowledgments

The MIAME document is a result of the work of many people. The idea was conceived during an international meeting organized by the EBI in November 1999 to discuss gene expression databases[8], during which a preliminary version of the MIAME document was produced. A microarray annotations mailing list was created and many of the members of this mailing list contributed to subsequent drafts. We would particularly like to acknowledge G. Barton, K. Henrick and J-J. Riethoven, M. Bittner, R. Bumgarner, M. Cherry, T. Freeman, J. Hoheisel and his team, A. Lash, H. Mangalam, T. Preiss, A. Richter, C. Schwager, M. Ringwald, Y. Tateno and R. Young. The document was extensively discussed in a meeting of the MGED steering committee meeting at the US National Institutes of Health in November 2000, where the current version of MIAME was effectively prepared. The final additions to the document were made during the MGED 3 conference[1] (http://www.mged.org/). Although the MGED is a grass-roots movement and does not presently have a dedicated funding, the authors of this paper have been funded from contributions from various sources, including the Industry Support Programme at the EBI, Lipper Foundation, Medical Research Council, Incyte Genomics and the National Heart, Lung, and Blood Institute of the US NIH.

Received 13 July; accepted 22 October 2001.

1. The Chipping Forecast. *Nature Genet.* **21**, 1–60 (1999).
2. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21**, 33–37 (1999).
3. Young, R. Biomedical discovery with DNA arrays. *Cell* **102**, 9–16 (2000).
4. Lockhart, D. & Winzeler, E. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
5. Aach, J., Rindone, W. & Church, G.M. Systematic management and analysis of yeast gene expression data. *Genome Res.* **10**, 431–445 (2000).
6. Quackenbush, J. Computational analysis of microarray data, *Nature Rev. Genet.* **2**, 418–427 (2001).
7. Sherlock, G. *et al.* The Stanford Microarray Database. *Nucleic Acids Res.* **29**, 152–155 (2001).
8. Brazma, A., Robinson, A., Cameron, G. & Ashburner M. One-stop shop for microarray data. *Nature* **403**, 699–700 (2000).
9. Editorial. Free and public expression. *Nature* **410**, 851 (2001).
10. Bassett, D.E., Eisen, M.B. & Boguski, M.B. Gene expression informatics—it's all in your mine. *Nature Genet.* **21**, 51–55 (1999).
11. Kerr, M.K. & Churchill G.A. Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201 (2001).
12. The Governing Council of the Organization for Human Brain Mapping (OHBM). Neuroimaging databases. *Science* **292**, 1673–1676 (2001).