SGMD: the Soybean Genomics and Microarray Database

Nadim W. Alkharouf1,2 and Benjamin F. Matthews1,*

1Soybean Genomics and Improvement Laboratory, USDA-ARS, Building 006, Room 118, 10300 Baltimore Avenue, Beltsville, MD 20705-2350, USA and 2School of Computational Sciences, George Mason University, 10900 University Boulevard, MSN 5B3, Manassas, VA 20110, USA

Received August 14, 2003; Revised October 15, 2003; Accepted October 23, 2003

ABSTRACT

The Soybean Genomics and Microarray Database (SGMD) attempts to provide an integrated view of the interaction of soybean with the soybean cyst nematode and contains genomic, EST and microarray data with embedded analytical tools allowing correlation of soybean ESTs with their gene expression profiles. SGMD provides analytical tools to mine the microarray data quickly by integrating many analysis methods within the database itself. The expression profiles of genes at time intervals during the first 8 days of nematode invasion is searchable by gene name or GenBank accession number. Recent developments include the addition of a searchable database for soybean cyst nematode ESTs and photographs of the invasion process at time points examined using microarrays. SGMD is completely accessible from the web at: http://psi081.ba.ars.usda.gov/SGMD/default.htm.

INTRODUCTION

Soybean (Glycine max L. Merr.) is the most important grain legume crop grown in the United States. The estimated value of soybean is 10.6 billion dollars. There are over 60 million acres planted with an average production of 32.4 bushels (1.14 cubic meter) per acre. The soybean cyst nematode (SCN), Heterodera glycines Ichinoe, is the major pest of soybean, causing an estimated $1 billion in damages throughout the USA per annum. This is more than the soybean loss from all other pests combined (1). The soybean genome is not fully sequenced yet, and little is known about the functions of many soybean genes, especially related to disease and pest resistance. The Soybean Genomics and Microarray Database (SGMD; http://psi081.ba.ars.usda.gov/SGMD/default.htm) is a public database that links soybean ESTs with gene expression data and provides embedded analytical tools for data mining. SGMD was established in 1999 to serve as a sequence and microarray database for the soybean community with its primary focus to store and analyze EST and microarray data generated from experiments involving the interaction of soybean with the SCN. The database currently stores over 50 million rows of microarray data (from USDA labs and collaborators) and almost 20 000 ESTs. Many of the ESTs are printed on microarray slides, allowing the correlation of expression levels with function. More recently, SGMD has broadened its scope to include public soybean ESTs and SCN ESTs. SGMD also serves as a means of novel gene discovery through EST analysis (2). Numerous applications to conduct statistical analysis on DNA microarray data have been integrated into the SGMD interface for rapid data analysis. Analytical tools include analysis of variance (ANOVA) and t-tests, which have been integrated using SQL procedures, thereby eliminating the need for third-party software.

DATABASE COMPONENTS

SGMD is a relational database built on SQLServer2000 and is housed at the Beltsville Agricultural Research Center in Beltsville, MD, USA. The web interface to SGMD is divided into three categories: microarray experiments, the sequence database and collaborations.

Microarray experiments

SGMD’s schema conforms to the minimal information about a microarray experiment (MIAME) guidelines set forth by the Microarray and Gene Expression Databases (MGED) group (3,4). MIAME aims to outline the minimum information required to unambiguously interpret microarray data and to subsequently allow independent verification of these data at a later stage if required. Data for a number of microarray experiments are accessible, among them a time series experiment examining soybean gene expression profiles during SCN invasion from 6 h to 8 days.

A number of web-based applications have been built, among them ANOVA and t-test applications that perform these tests to determine reproducibility and significance of measurements, respectively (Fig. 1). This is done on-the-fly from any web browser, but it works best using Internet Explorer. Other web-based tools allow the user to perform binary queries to identify commonly induced genes between two or more experiments. Cluster analysis data are also stored and accessible from the website (Fig. 2).
Sequence database

The sequence database provides access to EST sequences stored at SGMD. These include soybean ESTs sequenced at the USDA, Clemson University and Genome Solutions, Inc (GSI). All relevant information about every EST is stored, including the cloning vector and bacterial host strain, insert size, dbEST ID and GenBank accession number, type of clone (cDNA or genomic) and Blast results, which include E-value, score and identity percentage (5) (Fig. 3). Perl scripts were written to extract this information from the Blast results and are available through the authors. The database also contains results of EST analysis and contig assembly for specific GenBank root libraries.

Since the soybean genome is not fully sequenced yet, EST sequences such as the ones stored at SGMD will play an important role in gene identification and discovery, as they have in other organisms (6–8).

Collaborations

SGMD stores a number of microarray experiments that investigate issues other than the soybean–SCN interaction.
Among them are microarray experiments designed to detect genes important to the retention of the flat worm *Ascaris suum* in the intestine of swine (9) and to study the expression of genes in chicken upon infection by coccidiosis, an important disease of poultry (10).

**DATABASE ACCESS**

SGMD is accessible at http://psi081.ba.ars.usda.gov/SGMD/Default.htm. Raw data from microarrays can be downloaded and the manager of the database can be contacted by email at nalkhar3@gmu.edu.

**FUTURE PERSPECTIVES**

Work is underway to include pathway information, time series data analysis applications, and pictures documenting cytological events correlating with microarray profiling time points. Also, work is underway to implement MAGE-OM (Microarray Gene Expression–Object Model) and MAGE-ML (Microarray Gene Expression–Markup language), the data exchange model and the data exchange format, respectively (11), as set forth by MGED.

**ACKNOWLEDGEMENTS**

We thank Rana Khan, Imed Chouikha, Peggy MacDonald, Hunter Beard and Kris Pilitt for their technical expertise in generating the microarray and sequence data. This work is supported by the United Soybean Board (USB) and the CSREES-National Research Initiative (NRI) competitive grants program (grant no. 99-35302-8189). N.A. is supported by a pre-doctoral fellowship from the Beltsville Agricultural Research Center (BARC).

**REFERENCES**