The *Medicago* Genome Initiative: a model legume database

Callum J. Bell, Richard A. Dixon[1], Andrew D. Farmer, Raul Flores, Jeff Inman, Robert A. Gonzales[1], Maria J. Harrison[1], Nancy L. Paiva[1], Angela D. Scott[1], Jennifer W. Weller[2] and Gregory D. May[1],*

The National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, NM 87505, USA, [1]Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73402, USA and [2]Virginia Bioinformatics Institute, 1750 Kraft Drive Suite 1400, Virginia Tech, Blacksburg, VA 24061, USA

Received September 5, 2000; Revised and Accepted October 30, 2000

ABSTRACT

The Medicago Genome Initiative (MGI) is a database of EST sequences of the model legume *Medicago truncatula*. The database is available to the public and has resulted from a collaborative research effort between the Samuel Roberts Noble Foundation and the National Center for Genome Resources to investigate the genome of *M.truncatula*. MGI is part of the greater integrated *Medicago* functional genomics program at the Noble Foundation (http://www.noble.org), which is taking a global approach in studying the genetic and biochemical events associated with the growth, development and environmental interactions of this model legume. Our approach will include: large-scale EST sequencing, gene expression profiling, the generation of *M.truncatula* activation-tagged and promoter trap insertion mutants, high-throughput metabolic profiling, and proteome studies. These multidisciplinary information pools will be interfaced with one another to provide scientists with an integrated, holistic set of tools to address fundamental questions pertaining to legume biology. The public interface to the MGI database can be accessed at http://www.ncgr.org/research/mgi.

INTRODUCTION

Medicago truncatula is closely related to an important forage legume, alfalfa, and has been chosen as a model species for genomic studies in view of its small, diploid genome, short generation time, self-fertility and high transformation efficiency (1–3). Genes from *M.truncatula* share very high sequence identity to their counterparts from alfalfa (e.g. 98.7 and 99.1% at the amino acid levels for isoflavone reductase, and vestitone reductase, respectively), so it serves as an excellent, genetically tractable model for alfalfa which is tetraploid. In addition to alfalfa, *M.truncatula* can serve as a model organism for soybean and other economically important legumes.

As a legume, and unlike the most studied genetic model plant, *Arabidopsis*, *M.truncatula* establishes symbiotic relationships with nitrogen fixing Rhizobia. Roots of *M.truncatula* are also colonized by beneficial arbuscular mycorrhizal fungi (4). Furthermore, the complex interactions of legumes with microorganisms have resulted in the evolution of a rich variety of natural product biosynthetic pathways impacting both mutualistic and disease/defense interactions. Of these, the isoflavonoid pathway, which is not present in *Arabidopsis*, leads to nodulation gene inducers and repressors, pterocarpan phytoalexins involved in host disease resistance, and isoflavones with anticancer and other health-promoting effects for humans. This pathway has been well characterized in alfalfa, and in other legumes such as soybean and chickpea, at the metabolic, enzymatic and genetic levels (5). In addition, alfalfa has a rich and diverse complement of triterpene species, compounds that impact forage quality (6) and may serve as lead compounds for novel pharmaceuticals.

Medicago truncatula is currently the subject of major genomics initiatives. In the United States, an NSF-funded program coordinated by Doug Cook is producing ESTs (expressed sequence tags) and performing map-based cloning of symbiotic genes, comparative genomics and BAC survey sequencing (http://chrysie.tamu.edu/medicago/).

The Medicago Genome Initiative (MGI) is a database of EST sequences of the model legume *M.truncatula*. The database is available to the public and results from a collaborative research effort between the Samuel Roberts Noble Foundation and the National Center for Genome Resources to investigate the genome of *M.truncatula*. MGI's software system consists of three interacting sub-systems, a relational database for storage of the sequence data and the results of its analysis, an automated analysis pipeline that performs the analyses, and a user interface, which presents a variety of views of the data to the researcher. The user interface can be modified and developed somewhat independently of the other sub-systems. This allows considerable flexibility in implementing novel ways of analyzing and presenting data in response to the needs of the user community.

*To whom correspondence should be addressed. Tel. +1 580 221 7391; Fax: +1 580 221 7380; Email: gdmay@noble.org

RELATIONAL DATABASE OVERVIEW

The relational database supporting the MGI system is based upon a schema adapted from the one used by the Phytophthora Genome Initiative (PGI) (7). That schema was itself a descendent of the Genome Sequence DataBase (GSDB) schema (http://www.ncgr.org/research/sequence), which has proved to be a robust mechanism for storing DNA sequence information, and its associated meta data. Eight tables are used, the relations among them reflect the processing of the data by the Analysis Pipeline. Raw sequences and quality scores, as they arrive from the sequencing laboratory, are kept separate from sequences that have been screened for vector sequences and subjected to quality control. An action table stores the operations that are to be performed on each sequence, an analysis table specifies the nature of each analysis task and the specific parameters used, and a feature table stores fine-grained information from individual analysis tasks, such as the description line of any BLAST (8) similarities that are found. The schema upon which the PGI database is based has previously been described (7). The MGI schema is virtually identical to this, except for the addition of storage of sequence quality scores, the ability to distinguish between pre-release and already-published data, and the storage of GenBank accession numbers following public release. Both databases support the analysis pipeline by specifying the data upon which each stage operates, and storing information about tasks that need to be performed at each stage.

ANALYSIS PIPELINE

The sequence analysis pipeline is comprised of an ordered series of processing steps, each of which carries out a well-understood operation on its input, and creates output that may be used by a subsequent pipeline stage. For example, the vector-screening program uses raw sequencing data as its input, and produces truncated sequence that is then used by similarity searching algorithms. The execution of the pipeline is controlled by data that is itself housed in the database. Object-oriented software objects have been written which interact generically with these tables, in such a way that subclasses of these objects can be developed to implement many types of pipeline stages.

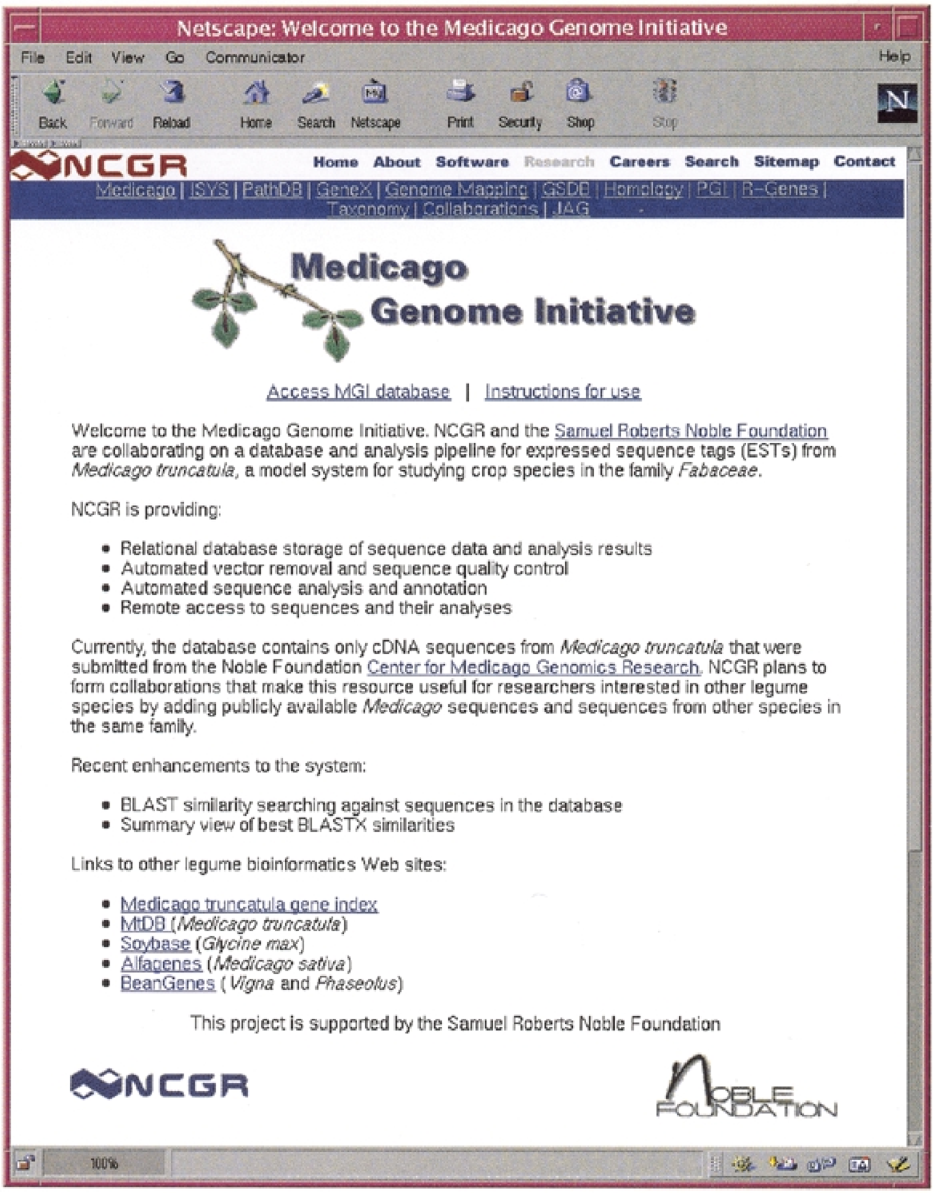
The MGI pipeline begins with the deposit of raw sequencing data and quality scores into a secure FTP site at NCGR. For each sequence file, a quality file is also deposited, consisting of a space-delimited array of numerical scores. The sequences and quality scores are outputs of the TraceTuner (Paracel) base-calling software. The PGI pipeline (7) is currently implemented at NCGR using a more simple architecture than the one used by MGI, but the processing steps are essentially the same. Accordingly, the common features will be covered only in outline. The pipeline is controlled by command scripts that invoke the specific pipeline stages in order and pass them relevant parameters. The scripts are scheduled as ‘cron’ jobs, and the MGI pipeline is run nightly.

A major difference between the prototype PGI system and MGI is the deployment in the latter of NCGR’s Distributed Analysis (DA) server to carry out computationally intensive tasks. The DA Server is a general facility that receives requests from bioinformatics software clients, and places them on a common queue, from which jobs are dispatched to a 40-processor domain of NCGR’s Sun Enterprise 10000. The system can be reconfigured to work with a different number of processors, or on a network cluster of machines. The jobs on the queue are processed in FIFO (first in, first out) order. Currently the DA server only serves computation in the BLAST family of programs, dividing the 40-processor domain into 10, 4-processor sub-domains, and running BLAST searches with four threads each. We plan to extend the type of computations that are served by the system to include other types of analyses such as sequence clustering, multiple alignment, and motif searching. Communication between the DA server and its clients, and between the server’s distributed elements, are supported with the Common Object Request Broker Architecture (CORBA) (9). The main benefit of the DA server is to reduce the time required for large sets of BLAST searches. As the database increases in size, a point will eventually be reached where analysis computation cannot keep up with growth—use of the DA server will delay this eventuality, and permit more frequent re-runs of important analysis tasks.

USER INTERFACE

MGI relies solely on WWW browsers as the remote user interface. The interface software consists of a static HTML launch page, and a set of CGI scripts written in the Perl programming language. These scripts execute pre-defined SQL (Structured Query Language) queries on the database, manipulate the data and present it to the user in a variety of ways. The use of Perl/CGI to support the user front-end has several advantages. The ways of querying and displaying the data are limited only by the possible complexity of database queries and the ability of HTML to display the data. Users are welcomed by a static HTML welcome page (Fig. 1), which has links to help pages, and to an analysis page that offers multiple options to access the data (Fig. 2). The welcome page has a link to a Perl CGI script that generates a web page with four choices of how to access data. Each of the subsequent choices invokes a Perl CGI script that executes a query on database via Sybperl or Perl DBI interfaces, and returns data to the main program which presents it to the user via their WWW browser. Java script is used to manage the browser window environment, creating new windows where appropriate.

Researchers are given a comprehensive choice of ways of looking at their data. Sequences can be selected by name, which produces a display that distinguishes the regions identified by the vector-screen and QC programs. Sequences can be saved to the user’s local disk in FASTA or plain format. Hyperlinked views of the BLASTN and BLASTX search results are available, along with a means of searching these results with keywords. Additionally, users can perform their own BLASTN, TBLASTN or TBLASTX searches against the MGI database itself, and view the results as hyperlinked HTML output or have them sent to their email address. Finally, the best BLASTX hits for each sequence can be viewed in a table that summarizes the salient features of each hit: the species of origin, the GenBank accession of the best hit (as a hyperlink to NCBI), the score and the Expectation value. The public interface to the MGI database is http://www.ncgr.org/research/mgi.


Netscape: Welcome to the Medicago Genome Initiative
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
NCGR
Home About Software Research Careers Search Sitemap Contact
Medicago | ISYS | PathDB | GeneX | Genome Mapping | GSDB | Homology | PGI | R-Genes | Taxonomy | Collaborations | JAG
100%

Medicago Genome Initiative

Access MGI database | Instructions for use

Welcome to the Medicago Genome Initiative. NCGR and the Samuel Roberts Noble Foundation are collaborating on a database and analysis pipeline for expressed sequence tags (ESTs) from *Medicago truncatula*, a model system for studying crop species in the family *Fabaceae*.

NCGR is providing:

- Relational database storage of sequence data and analysis results
- Automated vector removal and sequence quality control
- Automated sequence analysis and annotation
- Remote access to sequences and their analyses

Currently, the database contains only cDNA sequences from *Medicago truncatula* that were submitted from the Noble Foundation Center for Medicago Genomics Research. NCGR plans to form collaborations that make this resource useful for researchers interested in other legume species by adding publicly available *Medicago* sequences and sequences from other species in the same family.

Recent enhancements to the system:

- BLAST similarity searching against sequences in the database
- Summary view of best BLASTX similarities

Links to other legume bioinformatics Web sites:

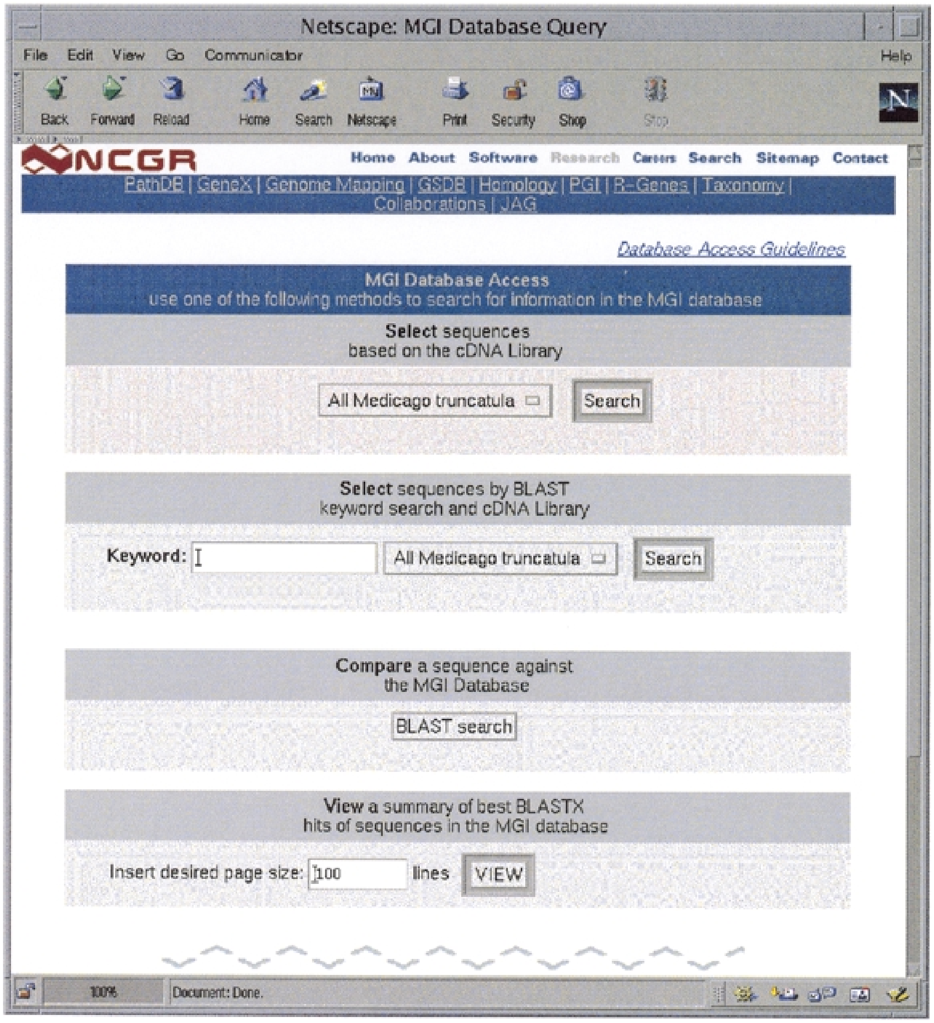
- Medicago truncatula gene index
- MtDB (*Medicago truncatula*)
- Soybase (*Glycine max*)
- Alfagenes (*Medicago sativa*)
- BeanGenes (*Vigna* and *Phaseolus*)

This project is supported by the Samuel Roberts Noble Foundation

Figure 1. The welcome page seen by MGI users in the public domain. Hyperlinks to help pages and to the main analysis page are provided.

CONCLUSIONS

We have developed a DNA analysis and database system for EST data. The system removes almost the entire bioinformatics burden from the DNA sequencing laboratory, other than the need for base calling and the assignment of quality scores to the data. Basing MGI on the prototype PGI system, we have added the ability to track whether a sequence has been submitted to GenBank, implemented a sequence quality-control algorithm based on PHRED or TraceTuner base quality scores, added the ability to query sequences based on their cDNA library of origin, implemented a sytem for users to query the database using the BLAST family of algorithms, and implemented a distributed analysis system that enables the use of workstation clusters or multiprocessor computers for computationally intensive tasks. The MGI system offers researchers interested in legume biology a set of useful tools for investigating EST sequences from their model system, and the site also serves as a starting point for further web-based bioinformatics exploration of species in the family Fabaceae, with links to other sites of importance. An attractive feature of the system is its independence of the organism being studied, and the relative ease with which it can be applied to new projects. Creating and configuring a database, pipeline and user interface for a new collaboration is relatively straightforward. The system has certain limitations. It is highly focused on individual ESTs as the units for analysis, rather than consensus sequences derived from EST clusters, and has no capacity for the analysis of genomic DNA. Clustering and multiple alignment of the ESTs are routinely performed, but this needs to become part of the analysis pipeline. We are presently designing the next generation of this architecture to overcome these limitations, add analysis methods, make the


Netscape: MGI Database Query
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
NCGR
Home About Software Research Careers Search Sitemap Contact
PathDB | GeneX | Genome Mapping | GSDB | Homology | PGI | R-Genes | Taxonomy | Collaborations | JAG
Database Access Guidelines
Search
Search
BLAST search
VIEW
100%
Document: Done.

MGI Database Access
use one of the following methods to search for information in the MGI database

Select sequences
based on the cDNA Library

All Medicago truncatula

Select sequences by BLAST
keyword search and cDNA Library

Keyword:

All Medicago truncatula

Compare a sequence against
the MGI Database

View a summary of best BLASTX
hits of sequences in the MGI database

Insert desired page size: 100 lines

Figure 2. The main analysis page. Four options are available for visualizing sequences and the results of their analysis, each initiated by interacting with the features on a separate row of the main table.

pipeline operation seamless, and to simplify the establishment and operation of new pipelines.

ACKNOWLEDGEMENTS

We gratefully acknowledge Peter Hraber, Bruno Sobral and the Phytophothora Consortium, upon whose database and analysis system, the Phytophothora Genome Initiative (PGI), this work was built. This project was supported by the Samuel Roberts Noble Foundation.

REFERENCES

1. Barker,D.G., Bianchi,S., Blondon,F., Dattée,Y., Duc,G., Essad,S., Flament,P., Gallusci,P., Génier,G., Guy,P., Muel,X., Tourneur,J., Dénarié,J. and Huguet,T. (1990) Medicago truncatula, a model plant for studying the molecular genetics of the Rhizobium-legume symbiosis. *Plant Mol. Biol. Reporter*, **8**, 40–49.
2. Cook,D.R. (1999) Medicago truncatula – a model in the making! *Curr. Opin. Plant Biol.*, **2**, 301–304.
3. Trieu,A.T., Burleigh,S.H., Kardailsky,I.V., Maldonado-Mendoza,I.E., Versaw,W.K., Blaylock,L.A., Shin,H., Chiou,T.-J., Katagi,H., Dewbre,G.R., Weigel,D. and Harrison,M.J. (2000) Transformation of *Medicago truncatula* via infiltration of seedlings or flowering plants with *Agrobacterium. Plant J.*, **22**, 531–542.
4. Harrison,M.J. and Dixon,R.A. (1993) Isoflavonoid accumulation and expression of defense gene transcripts during the establishment of vesicular arbuscular mycorrhizal associations in roots of *Medicago truncatula. Mol. Plant Microbe Interact.*, **6**, 643–654.
5. Dixon,R.A. (1999) Isoflavonoids: Biochemistry, Molecular Biology, and Biological Functions. In Sankawa,U. (ed.), *Comprehensive Natural Products Chemistry*. Elsevier, Oxford, UK, Vol. 1, pp. 773–823.
6. Small,E. (1996) Adaptations to herbivory in alfalfa (*Medicago sativa*). *Can. J. Bot.*, **74**, 807–822.
7. Waugh,M., Hraber,P., Weller,J., Wu,Y., Chen,G., Inman,J., Kiphart,D. and Sobral,B. (2000) The Phytophthora Genome Initiative database: informatics and analysis for distributed pathogenomic research. *Nucleic Acids Res.*, **28**, 87–90.
8. Altschul,S.F., Madden,T.L, Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
9. The Object Management Group, Inc. (1997) *The Common Object Request Broker: Architecture and Specification*, rev 2.1. The Object Management Group, Inc., Framingham, MA.