D660–D665 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki128

The Legume Information System (LIS): an integrated information resource for comparative legume biology

Michael D. Gonzales, Eric Archuleta, Andrew Farmer, Kamal Gajendran, David Grant[1], Randy Shoemaker[1], William D. Beavis* and Mark E. Waugh[2]

National Center for Genome Resources, Santa Fe, NM 87505, USA, [1]USDA-ARS-CICGR and Department of Agronomy, Iowa State University, Ames, IA 50011, USA and [2]Bio5, University of Arizona, Tucson, AR 85721, USA

Received August 15, 2004; Revised and Accepted October 24, 2004

ABSTRACT

The Legume Information System (LIS) (http://www.comparative-legumes.org), developed by the National Center for Genome Resources in cooperation with the USDA Agricultural Research Service (ARS), is a comparative legume resource that integrates genetic and molecular data from multiple legume species enabling cross-species genomic and transcript comparisons. The LIS virtual plant interface allows simplified and intuitive navigation of transcript data from *Medicago truncatula*, *Lotus japonicus*, *Glycine max* and *Arabidopsis thaliana*. Transcript libraries are represented as images of plant organs in different developmental stages, which are selected to query the analyzed and annotated data. Complex queries can be accomplished by adding modifiers, keywords and sequence names. The LIS also contains annotated genomic data featuring transcript alignments to validate gene predictions as well as motif and similarity analyses. The genomic browser supports comparative analysis via novel dynamic functional annotation comparisons. CMap, developed as part of the GMOD project (http://www.gmod.org/cmap/index.shtml), has been incorporated to support comparative analyses of community linkage and physical map data. LIS is being expanded to incorporate gene expression and biochemical pathways which will be seamlessly integrated forming a knowledge discovery framework.

INTRODUCTION

Legumes (soybeans, dry beans, peas, etc.) are excellent vegetable sources for proteins and oils, but are also invaluable as organic fertilizers because of their ability to fix atmospheric nitrogen. For these reasons, as well as their global economic importance, legumes have become the focus of increased interest and research activity. Recently, an international effort to sequence the gene-rich regions in *Medicago truncatula* has begun (http://www.medicago.org/genome/). *M.truncatula*, like many model organisms, is genetically tractable but has little agro-economic significance. In contrast, *Glycine max* (soybean) is grown on every major continent as a major vegetable source for protein and is a multi-billion dollar/year crop. Owing to its economic importance there has been considerable effort devoted to understanding the genetic basis for a number of economically important traits; however, the soybean genome is an ancient polyploid and is not likely to be completely sequenced in the near future. Fortunately, owing to the conservation of gene structure and function among related plant species, it is possible to leverage information through comparative genomics from model plants, such as *M.truncatula*, to better understand the relationship between genotype and phenotype in crop species.

Toward this goal the National Center for Genome Resources (NCGR), in cooperation with the USDA Agricultural Research Service (ARS), has developed the Legume Information System (LIS) (http://www.comparative-legumes.org). The LIS is a publicly accessible legume information resource that integrates genetic and molecular data from multiple legume species and enables cross-species genomic and transcript comparisons.

The intent of LIS is not to duplicate related efforts (http://www.comparative-legumes.org/lis/lis_links.html) but to leverage data-rich model plants to fill knowledge gaps across crop plant species and provide the ability to traverse between interrelated data types.

LIS ARCHITECTURE

LIS integrates map, genomic and transcript data from a number of sources and allows users to access and compare data via a single but multifaceted web interface (Figure 1). LIS has a

*To whom correspondence should be addressed. Tel: +1 505 995 4412; Fax: +1 505 995 4412; Email: wdb@ncgr.org

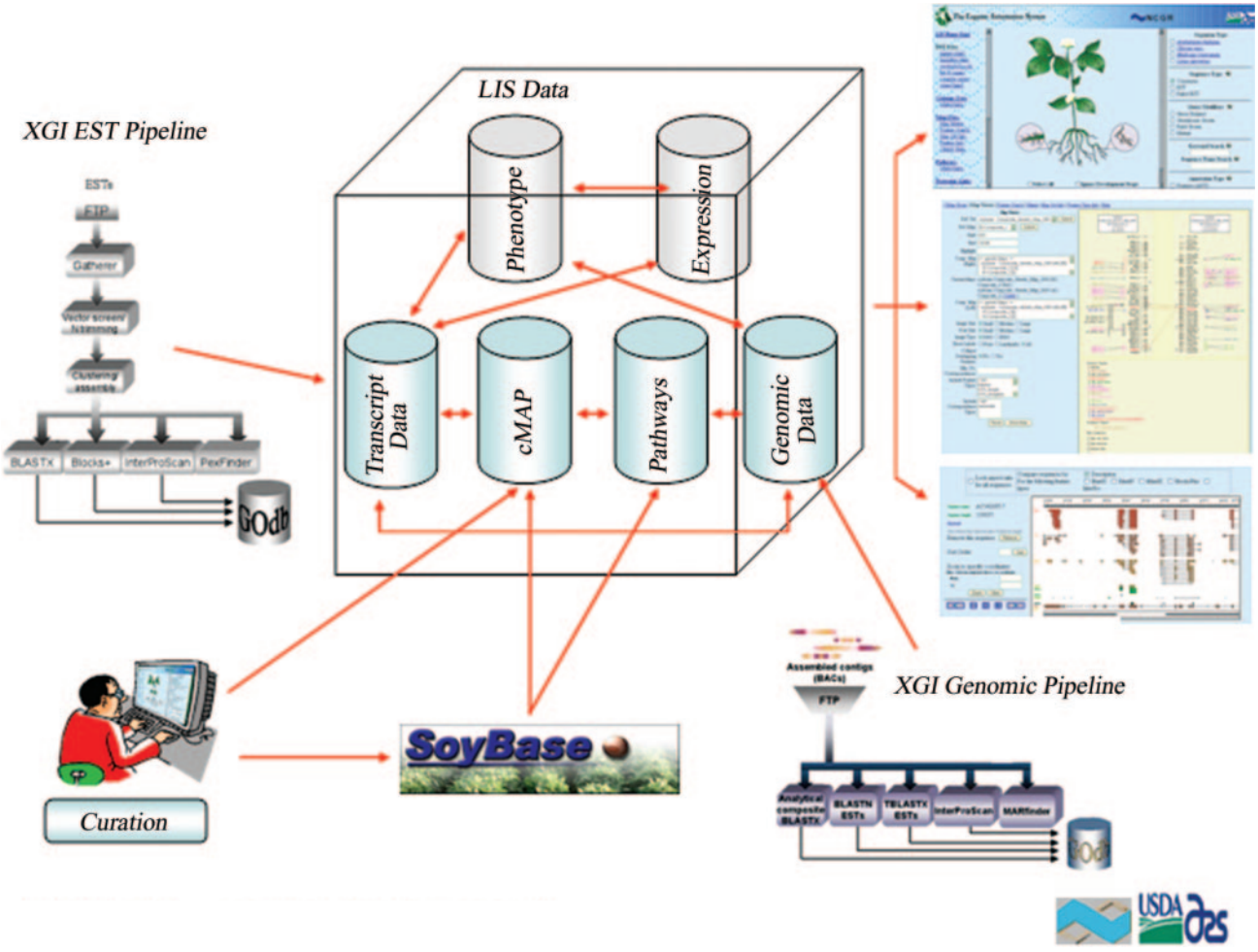

Figure 1. LIS architecture.

modular architecture, which can best be described as a local federation of databases delineated by data type. This allows each module to be updated independently reflecting the reality of differential data generation in the scientific community. For example, the genomic module is updated with phase 2 and 3 sequences on a near-continual basis, whereas the transcript module is updated based on EST count thresholds set for each species. A core database module has been developed to encapsulate data common to more than one module, for instance citations, which can reference genomic sequences within the context of physical maps or gene sequences in the context of pathways.

SoyBase (http://www.soybase.org), a resource familiar to the soybean community, has been partially integrated into the LIS, capturing all of the soybean map and biochemical pathway data. Additionally, all publicly available transcript and genomic data from *M.truncatula*, *Lotus japonicus*, *G.max* and *Arabidopsis thaliana* have been analyzed using NCGR's XGI (Genome Initiative for species X) system (http://www.ncgr.org/xgi). Analysis is performed within each species and the results can be used in cross-species comparisons to identify legume-specific gene sequences. Comparative transcript data are used in genomic analyses by aligning gene sequences to genomic contigs to validate *ab initio* gene predictions. CMap, developed by Ken Clark as part of the Generic Model Organism Database (GMOD) project (http://www.gmod.org/cmap/index.shtml), has been incorporated into LIS to provide comparative access to genetic and physical maps from a number of species. Future plans include integration of biochemical pathway information and gene expression data.

CONTENT

LIS makes use of the XGI system for automated analysis and annotation of both genomic and expressed sequence data. Sequence data are processed through a series of analysis operations, each operation building upon the results of the previous stage.

Genomic data

LIS genomic data are processed using the XGI genomic pipeline. Public sequence information is gathered from the NCBI's High Throughput Genomic (HTG) division (1) for the species of interest. HTG sequences are from large-scale genome sequencing centers and are submitted as in-process assemblies in various stages of completeness, often containing two or

more contigs; LIS does not assemble its genomic data, but takes data from GenBank as submitted by the sequencing centers. Each genomic sequence is then separated into its constituent contigs and analyzed in pieces using a sliding window of length 10 000 with an overlap of 3000 bp. These pieces are processed using BLASTX (1–3) against NCBI's nr database (1), and with BLASTN (2,3) and TBLASTX (2,3) against the consensus sequences produced in the LIS transcript database. BlimpSearcher analysis (4) against the Blocks+ database (5,6) is used to identify protein blocks. InterProScan (7) is run to integrate results from a variety of protein motif analysis tools using the InterPro database (8). Analysis results that are in common between overlapping pieces are merged before being stored. GenScan (9), which performs *ab initio* gene prediction on the genomic sequences, is run on the complete contig sequences, providing the opportunity to define and compare the genes and exon–intron organization of the sequences. The results of the genomic pipeline are stored in the LIS genomic database and are updated periodically depending on the number of sequences available for analysis.

Transcript data

The LIS transcript database consists of EST, Consensus and Failed EST sequences for *M.truncatula*, *G.max*, *L.japonicus* and *A.thaliana*. Using the XGI transcript pipeline, raw public EST and cDNA data are gathered from the NCBI and analyzed. Where available, quality scores for EST sequences are incorporated into the database for use by downstream analysis components. Detailed metadata concerning sequence origin, such as submitting organization, organism, library details and cloning methodology, are captured and are viewable in the LIS interface; libraries are also categorized by a manual curation process for more accurate querying through the interface.

Before being used in analyses, raw EST data are screened for quality. Screening operations include removal of most common vector sequences, poly(A/T) trimming, N-trimming, adapter/linker removal, length trimming and poor quality read trimming. Vector screening and adapter/linker screening removes sequence contamination of the insert that typically arises as part of the cloning process. In addition, the fidelity of a sequence read typically degenerates toward the end of the sequence, resulting in errors in base calling which are trimmed out as part of this process. Finally, low-complexity sequences represented by polyadenylated regions can produce many false positive matches in downstream analyses. The end result of the quality screens is a high-quality 'approved' sequence that is then deposited in the database. An EST that has failed the XGI vector screen analysis for one or more reasons is not included in subsequent analyses, but may still be inspected through the interface as a failed EST.

Approved EST data are clustered using Phrap (http://www.phrap.org), which performs clustering and contig assembly to produce Consensus sequences; these aggregate the high-quality sequence information of their member ESTs and are used in all downstream analyses. Consensus sequences are analyzed using NCBI's BLASTX algorithm (1–3) to search for potential homologs against NCBI's nr database (1). BlimpSearcher (4) and InterProScan (7) are used as previously described. Each of these analyses is followed by a stage that uses the results to associate Gene Ontology (GO) terms (10) with the sequences. Pexfinder (http://www.oardc.ohio-state.edu/phytophthora/pexfinder), co-developed by NCGR and the Kamoun laboratory at OSU-OARDC and based on Signal P (11) has also been incorporated. Pexfinder (Protein EXcreted) predicts proteins excreted through the plasma membrane based on signal peptides. The results of the pipeline analyses are stored in the LIS transcript database and are updated periodically depending on the number of public sequences available for analysis.

Map data

All curated linkage map data from SoyBase have been subsumed and incorporated into the CMap system (http://www.gmod.org/cmap/index.shtml), a publicly available comparative mapping tool developed by Ken Clark as part of the GMOD project (http://www.gmod.org). CMap provides LIS users access to curated *G.max* and *Phaseolus vulgaris* genetic maps and will soon incorporate genetic and physical maps for other species. Correspondences between markers on different maps have been pre-computed based on curated name matches, taking into account the possibility of the same marker being mapped onto multiple loci within a map set. As part of this project, CMap is being modified to support multiple comparators including sequence similarity in addition to marker name and curated relationships.

LIS USER INTERFACE

The LIS web interface provides a powerful set of search tools to access, compare and save transcript, genomic and map data. Sequence and analysis data are logically organized and searchable in a variety of ways. The LIS interface features a novel 'virtual plant' interface that allows simplified and intuitive navigation of all publicly available transcript data (Figure 2). Images of both mature and immature plants, as well as seeds and seed pods make it much easier for scientists to browse the analyzed and annotated transcript data simply by clicking on different parts of the virtual plant that represent different anatomical features or organs. More sophisticated queries using selected libraries, conditional modifiers, such as 'mutant' or 'greenhouse grown', and keyword searches of features and GO annotations are also supported. Searches can be restricted by organism, library and sequence type as well as by sequence name. The interface also gives users the ability to search for results and annotations based on specific types of analysis tool output (e.g. get only these sequences that have InterPro results) or combinations of output (retrieve only sequences with both Blocks+ and InterPro results). Precise delineation of library conditions can also be accomplished. For example, a researcher interested in stress-induced expression could choose to query for Consensus sequences from ANY of the stress-related libraries and NO others. This query would return Consensus sequences whose member ESTs only originate from the selected libraries, thereby eliminating constitutively expressed genes. This is performed using intuitive search options that use Boolean logic operators (AND, OR, NOT) in conjunction with scope delimiters (STRICT, LOOSE) thus enabling virtual northerns and *in silico* subtractions.

The interface presents data in a variety of formats, including graphical depictions of sequences decorated with their

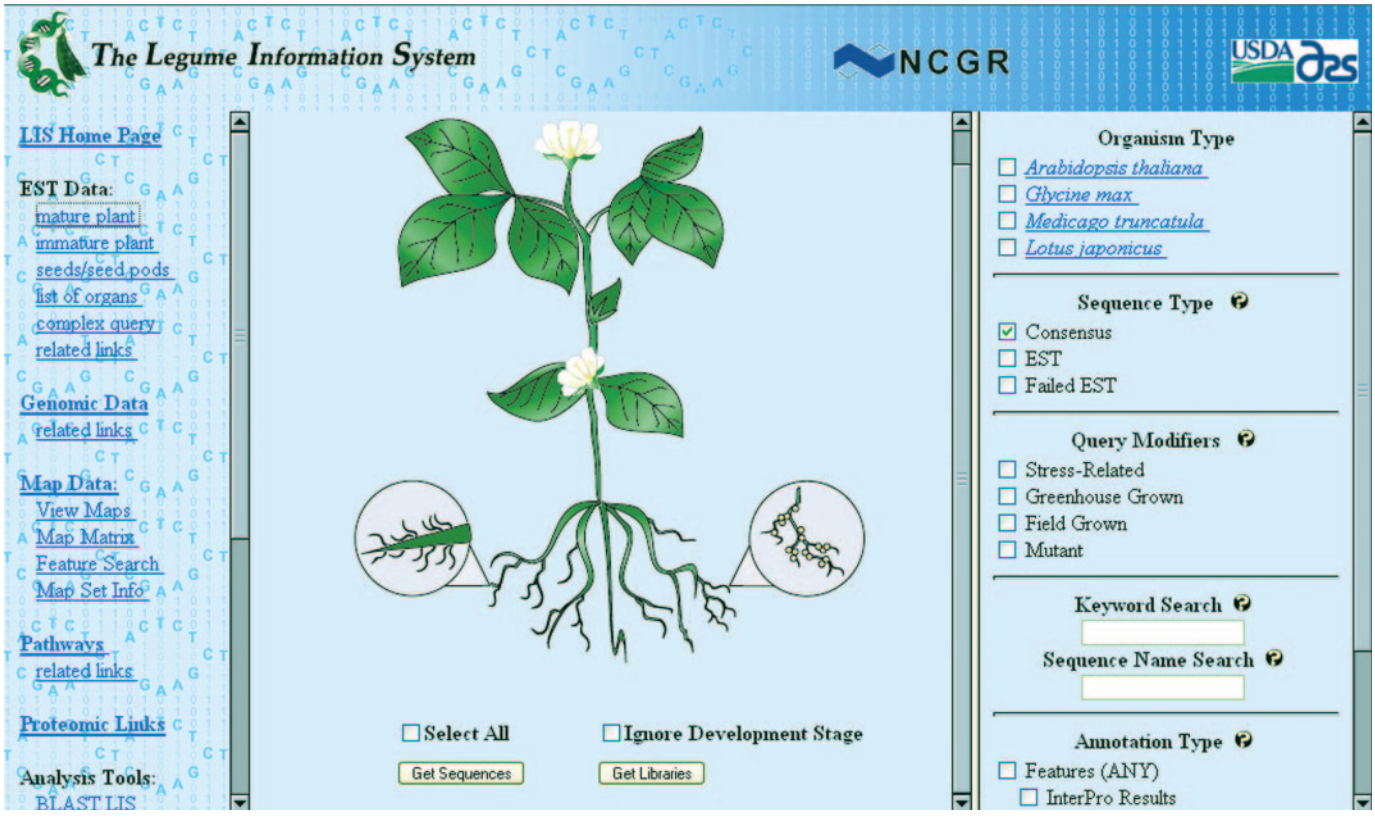

Figure 2. Clickable plant user interface.

predicted features, multiple sequence alignments with sequence variants highlighted and detailed reports of analysis operations.

Sequence details

The sequence detail views capture all information relevant at the individual sequence level, and other information can be accessed through this display. This includes quality trimming, base composition, as well as sequence metadata and clustering information. For Consensus sequences, it is also a portal to all analyses run on the sequence.

Features and annotations

The Features and Annotation (F&A) view displays all data available for Consensus sequences. Links to sequence details, multiple sequence alignments (MSAs), EST membership data as well as library and organism metadata are provided. The F&A also gives a graphical presentation of the analysis results linked directly to Gene Ontology annotations where appropriate. In addition, output from the analysis results can be viewed by following the appropriate links.

Multiple sequence alignments

Clustering results are summarized as MSAs. MSAs are associated with any sequence in the database (both Consensus sequences, and ESTs) with mismatches between EST and consensus sequences highlighted in red.

Creating a custom account

Users may create an LIS account by filling out some simple contact information. Registered users have the ability to save sequences and queries of interest to the MyData component of LIS, as well download data in a variety of formats. In the future, registered users may also elect to receive emails regarding news, events and updates to the LIS system. MyData is a tool for archiving and saving sequences and queries using personalized folder and query names. The MyData page supports bulk downloads of sequences in forward, reverse compliment or both orientations in the FASTA format as well as downloading sequence analysis and annotation details via the Summary Download option. Using the Summary Download feature, users may choose to download analysis results, EST quality screen details, cluster information, as well as feature location information for both genomic and transcript data in tab-delimited text or Excel formats. Access to user data is password protected: a temporary password is issued via email at the time of registration and can be changed when the user logs in. For the un-registered user, the interface supports individual sequence downloads in the FASTA format.

Sequence comparisons

Users have the ability to BLAST local sequences against LIS transcript, consensus and genomic sequence data by pasting one or more FASTA formatted sequences into a conventional BLAST interface or by browsing the users local hard drive for sequence files. Concatenated sequences may also be used provided they are in the FASTA format. Target database options

include individual species for both genomic and transcript as well as EST and Consensus datasets.

The Comparative Functional Genomics Browser

The Comparative Functional Genomics Browser (CFGB) visualizes genomic analysis results, including comparative transcript data aligned to genomic contigs to validate gene-predictions (Figure 3). The CFGB also supports serial addition of contigs to the browser for comparative alignments and novel dynamic functional annotation comparisons by using description matches of the different analysis types. Soon, the CFGB will support comparisons via shared GO terms. Each genomic sequence has been annotated using the XGI genomic pipeline and each colored block represents the analysis type as well as the location and directionality of a match in relation to the genomic sequence. The size and orientation of the images in the CFGB can be manipulated by the

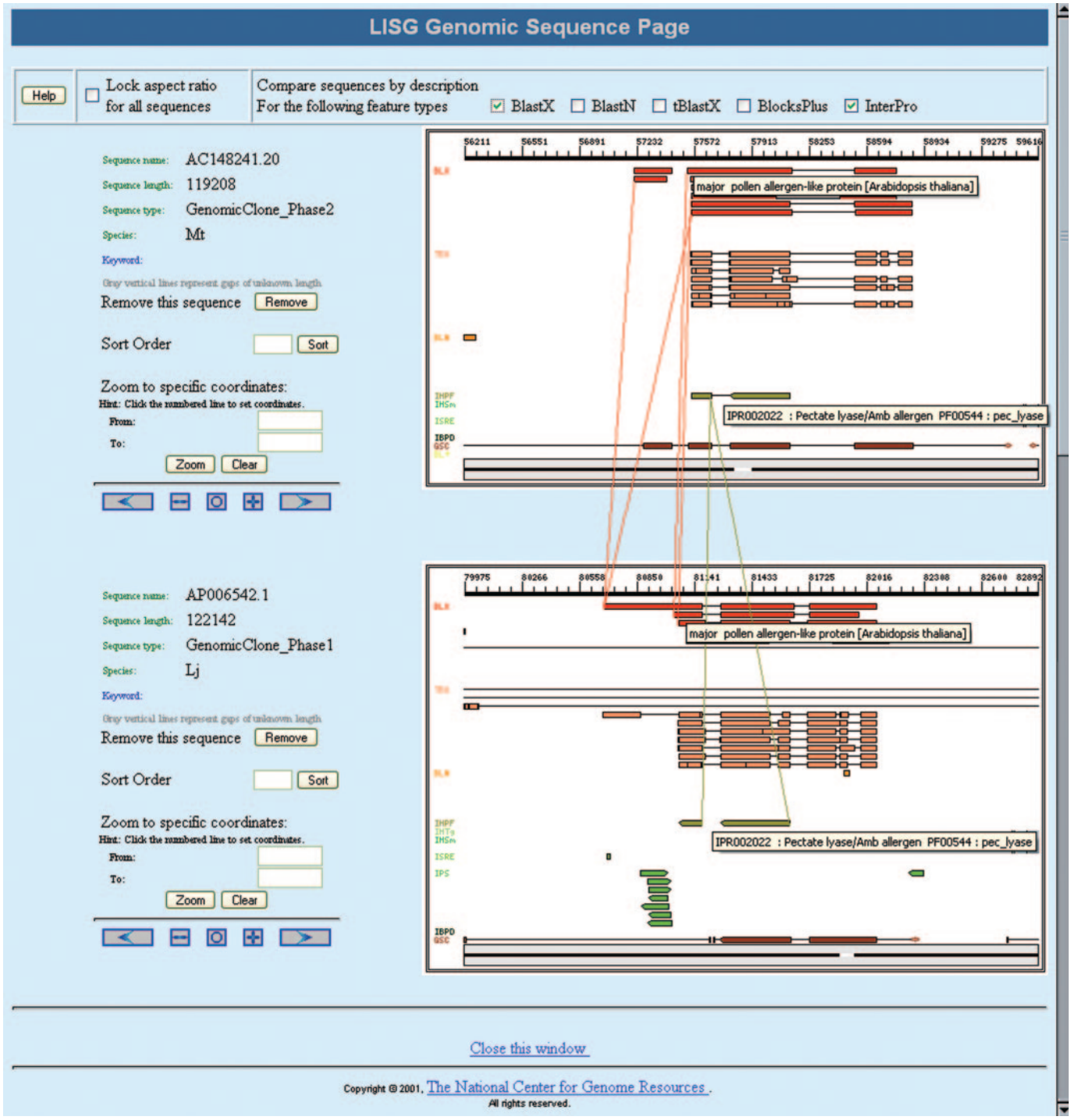

Figure 3. The Comparative Functional Genomics Browser.

zooming, panning and sorting functions. The CFGB also supports the ability to change the aspect ratios for all sequences at the same time. In the near future, the CFGB browser that is will be augmented by a synteny browser being developed by Volker Brendel's group at Iowa State University (http://gremlin1.zool.iastate.edu/~volker/). Unlike CFGB, which compares annotations between user-selected sequences on the fly, the synteny browser displays pre-computed relationships between contigs at the nucleotide and protein level. It is anticipated that the user may discover evidence of possible micro- or macro-syntenic relationships with CFGB which can be explored more fully using the synteny browser.

CMap interface

CMap features positioned on the maps are displayed using different symbols and colors to represent the various feature types; e.g. the QTL data from SoyBase is color coded according to a high-level classification of QTL types. When using the tool for comparative work, the user can choose any map for the database as a reference map in the given comparison. Comparative maps can then be added to the left or right of the reference map from a list of all maps in the database having a minimal set of correspondences to the reference map. Alternatively, a comparison matrix displays the number of correspondences between different maps and map sets in the database, and can be used to locate maps with many correspondences. These features allow users to compare maps both within and between map sets. Lines indicating relationships between features are drawn between the corresponding features on maps in a comparison. Feature details and map set information can be accessed from the map view. Map set details include species, map type, map units, curator remarks and the listing of the maps. Feature details include the feature name, feature type, aliases or synonyms, map position, cross references to other databases as well as correspondence details from other maps associated with the feature. The set of features and correspondences displayed on a map can also be restricted based on attributes such as feature types and correspondence types. Sizing and saving the map images is also supported.

FUTURE DEVELOPMENTS

In the next year and a half of development, LIS will be incorporating pathway and expression data as well as adding additional transcript and genomic data for an additional number of species. In collaboration with the international *Medicago* genome sequencing effort, LIS will also support third party analysis and annotation through a MOBY-enabled CFGB. The MOBY-enabled browser will be able to display user-selectable annotation tracks from each of the participating sequencing centers ensuring that the canonical annotation, should it exist for a region, is displayed along with transcript alignments and motifs.

A common complaint in the scientific community concerns the different Unigenes/Tentative Consensus (TC)/Consensus sequence sets generated for a given organism by different groups. There are advantages and disadvantages to each set, but there is currently no simple way to compare them. To this end, a dynamic BLAST browser is being developed to compare Unigenes/TCs/Consensus sequences from different participating institutions using user-defined criteria. This tool, being developed in collaboration with Volker Brendel's group at Iowa State, will display the results of user-specified comparisons in the graphical format, complete with member ESTs aligned to the parent contigs used in the comparison.

In the near future the ability to seamlessly traverse between map, genomic and transcript data will be implemented. A compelling example will be to 'project' LIS-analyzed Consensus sequences onto a genetic map with QTL information using BLAST hits to genomic sequences and marker associations to the BACs. If the supporting data exists, users will be able to follow a phenotype from a region on a genetic map to a collection of annotated gene sequences aligned to a genomic contig and, by comparison of functional annotation and syntenic relationships, to chromosomes of other species as well.

ACKNOWLEDGEMENTS

We would like to thank the LIS Steering Committee: Perry Cregan, David Grant, Greg May, Henry Nguyen, Randy Shoemaker, Cari Soderlund, Lila Vodkin and Nevin Young for their useful feedback as well as Lincoln Stein's group at CSHL for developing CMap and providing valuable insight into AceDB. Finally, we would like to thank Dr Pan and other members of Volker Brendel's group for their continuing efforts. This project is funded by USDA-ARS Specific Cooperative Agreement #3625-21000-038-01.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank. *Nucleic Acids Res.*, **32**, D23–D26.
2. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., 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.
4. Henikoff,S. and Henikoff,J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics*, **19**, 97–107.
5. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
6. Henikoff,S., Henikoff,J.G. and Pietrokovski,S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.
7. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
8. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
9. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
10. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
11. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.