Plant genome informatics: Evaluation and analysis of genomic DNA features involved in the transcriptional processing of protein coding genes

by

Shannon Dwayne Schlueter

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Volker Brendel, Co-major Professor
Randy C. Shoemaker, Co-major Professor
Xiaoqiu Huang
Thomas Peterson
Xun Gu

Iowa State University

Ames, Iowa

2006

UMI Number: 3243822

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Dedicated to:

Jessica, Harmon, and Elyse

My love, My pride, My inspiration

TABLE OF CONTENTS

CHAPTER 1. GENERAL INTRODUCTION

Introduction

Genome sequencing efforts in the plant kingdom have produced an astounding amount of sequencing information in the last decade. *Arabidopsis thaliana* (TAGI, 2000) and *Oryza sativa* (Goff et al., 2002; Yu et al., 2002) were the first major plant genomes to be sequenced. However, genome sequencing efforts are now underway in *Zea mays*, *Medicago truncatula*, *Lotus japonicus*, *Lycopersicon esculentum*, *Manihot esculenta*, *Populus trichocarpa*, and most recently *Glycine max*. These projects are amassing a tremendous sequence resource for investigation of a wide range of biological questions. The computational tools to both manage and interpret this data, however, are often difficult to apply to hypothesis driven research. Furthermore, many of these tools are designed for use with data from Human or other vertebrate sequencing projects and have limited effectiveness for plant genome analysis.

In order to provide a sound infrastructure for analyzing the various sources of genomic data available in the plant sciences and for interpreting the results of computational tools applied to this data, the xGDB resource has been created. The first implementation of the xGDB structure was applied to the model plant *Arabidopsis thaliana*. Spliced alignments of EST, cDNA and the annotations of the Arabidopsis genome were parsed and imported into a MySQL relational database. An elaborate web interface was designed for the database to allow users to browse the genome and query the database by sequence similarity, identifiers, or description

(http://www.plantgdb.org/AtGDB/). In general, the web interface is composed of three parts: the genomic context view, the query view, and the sequence view. The genomic context view allows users to browse a specific genomic region in the context of multiple annotation resources. The region graphic displays multiple sources of alignment information relative to one another. The query view allows users to view and interact with the results of a user query. Stored EST/cDNA alignments and annotated transcripts each have an individual page, the sequence view, which brings together sequence data, analysis tools, and related external links. The web interface efficiently presents the database entries on the fly and facilitates data access and utilization. This system has been used in the analysis of gene annotation quality, 5'- and 3'- UTRs, non-canonical splicing, U12-specific splicing, alternative splicing, abnormal intron and exon sizes, and conserved homologous sequences. (Zhu et al., 2003; Schlueter et al., 2003; Schlueter et al., submitted). This system was integrated into the PlantGDB framework (Dong et al., 2004; Dong et al., 2005) and now provides information and tools for *Arabidopsis*, rice, maize, *Medicago*, *Lotus*, *Populus*, tomato, soybean, *Brassica*, wheat, and sorghum. In addition, this system has been used in conjunction with the yrGATE gene-structure annotation tool (Wilkerson et al., 2006) to annotate homeologous soybean BAC sequences (Schlueter et al., 2006a; Schlueter et al., 2006b).

Albeit the current methods of gene structure annotation are vastly more accurate and provide more complete coverage than those employed less than five years ago, the support of annotations on a per-gene basis differs extraordinarily. Annotation quality can be directly attributed to the presence of expressed sequence alignments. The dependence of gene structure annotation on available EST and cDNA sequences makes static

assignments of gene structure problematic. For this reason, methods of analyzing and maintaining gene annotations must acknowledge confidence in a predicted gene structure. To provide these confidence estimators and an effective method for querying gene annotations based on these values, the Genome Annotation Evaluation Algorithm, GAEVAL, was developed (Schlueter et al., 2005; Schlueter et al, unpublished results).

Inconsistency of gene structure annotation is a limitation to research in the post-genome era. It is unrealistic to hope for better software solutions in the near future that will solve all or even a majority of the problems encountered by computational annotation tools. This issue is all the more urgent with an increasing number of species being sequenced and analyzed by comparative genomics – erroneous annotations could easily propagate. To address this limitation, a dynamic and economically feasible solution to the annotation predicament was developed by providing broad-based, web-technology-enabled community annotation tools (Schlueter et al., 2005; Wilkerson et al., 2006).

Previous understanding of the factors and sequence elements responsible for the initiation of eukaryotic gene transcription has been established primarily through conventional genetic analysis. However, these signals have been found to differ considerably among plants, animals and yeast. The majority of effort devoted to the *in silico* prediction of these regions and the analysis of their corresponding signals has been carried out in vertebrate species. As such, available programs that attempt to identify promoters by prediction of transcriptional start sites often perform poorly on plant sequences. Using a highly refined collection of sequences from Arabidopsis, a program

called TSiP was developed to predict plant promoter regions in anonymous DNA sequence.

Dissertation Organization

This dissertation is organized into five chapters. Chapter 1 contains a brief introduction to the areas of interest. Chapters 2 and 3 each consist of a published manuscript. Chapter 4 consists of a manuscript in preparation for publication. Chapter 5 is a summary of conclusions reached during the course of this dissertation research.

Chapter 2, entitled "xGDB: open-source computational infrastructure for the integrated evaluation and analysis of genome features" has been published in *Genome Biology* in 2006, Volume 7 electronic publication R111 Contributions to this work from co-authors include the following. Volker Brendel provided computational hardware and helpful discussion during the course of this work. Matt Wilkerson provided support in the development of this system and feedback during the writing of the manuscript. Qunfeng Dong has maintained the individual plant species xGDB instances as PlantGDB (http://www.plantgdb.org/) and has provided feedback during the writing of the manuscript. All of the code was developed by Shannon Dwayne Schlueter who was responsible for the writing of the manuscript.

Chapter 3, entitled "Community-based gene structure annotation" has been published in *Trends in Plant Science* in 2005, Volume 10 pages 9-14. Contributions to this work from co-authors include the following. Matthew Wilkerson contributed to the web development and documentation of this system. Volker Brendel provided use cases for the system and developed a curriculum for student involvement in science through the

use of this system. Eva Huala and Seung Y. Rhee provided the integration of this tool with the TAIR community database and the procedures for submission of annotations to the TAIR curation pipeline. Shannon Dwayne Schlueter developed and coded the GAEVAL system and was responsible for a majority of the writing.

Chapter 4, entitled "TSiP: transcriptional start site identification in plants" has been prepared for submission to *Genome Research*. The promoter prediction software described in this work, the creation of datasets used in developing and testing the predictive model reported by this work, and the writing of this manuscript are accomplishments of the sole author Shannon Dwayne Schlueter.

References

Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796-815.

Dong, Q., Schlueter, S.D., Brendel, V. 2004. PlantGDB, plant genome database and analysis tools. Nucleic Acids Res. 31: 3597-3600.

Dong, Q., Lawrence, C.J., Schlueter, S.D., Wilkerson, M.D., Kurtz, S., Lushbough, C., Brendel, V. 2005. Comparative plant genomics resources at PlantGDB. Plant Physiol. 139: 610-618.

Goff, S.A., Ricke D., Lan T.H., Presting G., Wang R., Dunn M., Glazebrook J., Sessions A., Oeller P., Varma H., Hadley D., Hutchison D., Martin C., Katagiri F., Lange B.M., Moughamer T., Xia Y., Budworth P., Zhong J., Miguel T., Paszkowski U., Zhang S., Colbert M., Sun W.L., Chen L., Cooper B., Park S., Wood T.C., Mao L., Quail P., Wing R., Dean R., Yu Y., Zharkikh A., Shen R., Sahasrabudhe S., Thomas A., Cannings R., Gutin A., Pruss D., Reid J., Tavtigian S., Mitchell J., Eldredge G., Scholl T., Miller R.M., Bhatnagar S., Adey N., Rubano T., Tusneem N., Robinson R., Feldhaus J., Macalma T., Oliphant A., Briggs S. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 296: 92-100.

Schlueter, J.A., Scheffler, B.E., Schlueter, S.D., Shoemaker, R.C. 2006a. Sequence conservation of homeologous BACs and transcription of homeologous genes in soybean (*Glycine max* L Merr). Genetics 174: 1017-1028.

Schlueter, J.A., Vasylendo-Sanders, I.F., Deshpande, S., Yi, J., Siegfried, M., Roe, B.A., Schlueter, S.D., Scheffler, B.E. Shoemaker, R.C. 2006b. The FAD2 gene family of soybean: insights into the structural and functional divergence of a paleopolyploid genome. The Plant Genome, in press.

Schlueter, S.D., Dong, Q., Brendel, V. 2003. GeneSeqer@PlantGDB: Gene structure prediction in plant genomes. Nucleic Acids Res. 31: 3597-3600.

Schlueter, S.D., Wilkerson, M.D., Huala, E., Rhee, S.Y., Brendel, V. 2005. Community-based gene structure annotation. Trends Plant Sci. 10: 9-14.

Schlueter, S.D., Wilkerson, M.D., Brendel, V. 2006. xGDB: open-source computational infrastructure for the integrated evaluation and analysis of genome features. Genome Biology 7: R111.

Schlueter, S.D. 2006. Tsip: transcriptional start site identification in plants. To be submitted to Genome Research.

Schlueter, S.D., Wilkerson, M.D., Brendel, V. 2006. The GAEVAL system: scoring eukaryotic gene structure annotations for authenticity based on EST and full-length cDNA evidence. Unpublished results.

Wilkerson, M.D., Schlueter, S.D., Brendel, V. 2006. yrGATE: a web-based gene-structure annotation tool for the identification and dissemination of eukaryotic genomes. Genome Biology 7:R58.

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W,

Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 296: 79-92.

Zhu, W., Schlueter, S.D., Brendel, V. 2003. Refined annotation of the Arabidopsis genome by complete expressed sequence tag mapping. Plant Physiol. 132: 469-484.

CHAPTER 2: XGDB: OPEN-SOURCE COMPUTATIONAL INFRASTRUCTURE FOR THE INTEGRATED EVALUATION AND ANALYSIS OF GENOME FEATURES

A paper published in *Genome Biology*[1]

Shannon D. Schlueter[2], Matthew D. Wilkerson[2], Qunfeng Dong[2], and Volker Brendel[2,3,4]

ABSTRACT

The eXtensible Genome Data Broker (xGDB) provides a software infrastructure consisting of integrated tools for the storage, display, and analysis of genome features in their genomic context. Common features include gene structure annotations, spliced alignments, mapping of repetitive sequences, and microarray probes, but the software supports inclusion of any property which can be associated with a genomic location. The xGDB distribution and user support utilities are available online at the xGDB project website http://xgdb.sourceforge.net/.

RATIONALE

Computational infrastructure is vital for all aspects of genome research. The assembled genomic sequence of an organism provides a natural scaffold for organizing biological

---

[1]Reprinted with permission of *Genome Biology,* 2006, 7, R111
[2]Department of Genetics, Development, and Cell Biology
[3]Department of Statistics, Iowa State University, Ames, IA 50011-3260, USA,
[4]Author for correspondence

data. However, researchers are easily overwhelmed without the computational tools necessary to interpret the features of these assemblies [1-4]. Although a large number of useful tools are available, they exist primarily as ad-hoc collections [5-7]. The xGDB software was designed to provide a framework for genomic data storage, display, and analysis and to provide integration of existing and novel genome analysis tools. The software is portable and easily installed for either public access or as a private workbench. It comes ready to use with the following capabilities:

- Detailed feature record pages
- Detailed views of genomic contexts
- Support for online community annotation
- Utilities for storage of feature data in relational databases
- Effortless integration and attachment of novel analysis tools
- Transcript View: A novel nucleotide resolution view of genomic contexts
- Compressed storage and dynamic retrieval of feature evidence alignments
- Attachment and organization of multiple URLs to any feature in any context
- Integrated heuristic searches based on feature identifier, alias, and/or description

It is important to note that xGDB differs from and is complementary to database systems like GMOD [8], EnsEMBL [9], or GenBank [10]. Unlike these systems which are tasked to provide encompassing data storage, xGDB instances are applied to specific research oriented tasks, which are enabled by the browser and integrated analysis tools. Because of the varying reliability of genomic features, there is a strong need to go beyond simply plotting such features for display (as would be available in GBrowse [8], for example) .

Contextual analysis of genomic features often requires filtering each feature by criteria specific to an individual user's needs. Such filtering requires the development of a system around a genome browser which manages storage and display of the evidence each feature is based on. Driven by this need, xGDB infrastructures provide interconnected analysis, visualization, and data management tools in a ready to use and easily extended package. The xGDB system is unique in providing this capability and is currently the only system to integrate Geneseqer [11] spliced alignment features.

An extensible infrastructure allows a wide array of data, tools, and analysis results to be brought together and provides the means by which to target their use in a focused manner. The xGDB package has been used to establish unique infrastructures tailored to the evaluation of genomic features. The xGDB instances available at PlantGDB [12] have been widely used in the analysis of genome annotation, gene structure determination, alternative splicing, and gene copy distribution [13-17]. Developing ad-hoc methods for such analyses is expensive and time-consuming. This cost is a major deterrent to many research endeavors and often leads to continuous redevelopment of analysis procedures [18-21]. Lack of stability leaves users questioning the accuracy of such analyses. The xGDB infrastructure provides both extensibility and procedural stability. Analysis procedures and results are made transparent to users allowing them to formulate their own opinion of results and providing a means to reproduce and maintain each analysis.

In the following we first discuss the features and capabilities of an xGDB system as seen by end-users. We then present the internal design and back-end components relevant to data providers and private installations. We should note that installation is

straightforward and requires only basic knowledge of commonly used open source software. For the purpose of illustration, we refer to AtGDB [22] and ZmGDB [23], publicly accessible xGDB instances established at PlantGDB. AtGDB and ZmGDB are respectively based on the five assembled chromosomes of *Arabidopsis thaliana* and emerging genomic sequence assemblies of *Zea mays*. Additional plant genome xGDB systems are accessible through the PlantGDB website [24].

FEATURES AND CAPABILITIES

The xGDB system is primarily accessed through dynamically generated web pages. These pages can be classified into context, record, and web service pages. Context pages present the location of genomic data sources in relation to surrounding features. Record pages localize pertinent external references, alignment results, and web service links. Web service pages allow a user to interact with data stored in the xGDB system, for example invoking BLAST for sequence comparisons [12, 25] or GeneSeqer for spliced alignment of transcript sequences [11, 26]. The whole set of web pages allows the system to quickly retrieve large amounts of data relevant to the user-specified task and control data presentation in a targeted and organized manner. By default, xGDB is configured to target data presentation for the purpose of evaluating gene structure annotation and genome annotation content, but has also been used to evaluate alternative splicing, microarray probe uniqueness, repetitive DNA positioning, and genetic marker placement.

Viewing genomic regions in context

Accessing an xGDB system, users are presented with navigational controls allowing them to search for genomic feature records and/or genomic locations. Navigational controls are displayed in a standard header at the top of all pages generated by the xGDB system (Fig.1.2). Depending on the configuration of xGDB, users may be presented with controls for selecting chromosomal coordinates from established genomic assemblies. These coordinates may be based on current or historic assembly versions, thus providing tracking of features occurring in previous assemblies. In lieu of chromosome based navigation, controls for selecting individual coordinate locations in smaller assemblies such as a single bacterial artificial chromosome (BAC) or genome survey sequence (GSS) may be provided. These controls fetch the genomic region spanning the user supplied coordinates and display a genomic context page.

Genome context pages contain one or more sources of feature data such as curated gene annotations, locations of genomic markers, alignments of microarray probes, gene structure predictions, and alignments of EST, cDNA or assembled contigs of sequence. Figure 1 contains a context display of ZmGDB including community contributed gene annotations, GenBank documented gene feature annotations, GSS alignments, alignments of homologous proteins, cDNA and EST alignments, the alignment of PlantGDB Unique Transcript (PUT) assemblies, and the alignment of microarray probes (Fig.1.7-14). Features may be represented by an assortment of glyph colors and shapes which can be used to visually distinguish properties specific to each. For example, in Figure 1 the context graphic showing EST alignment features (Fig.1.12) uses color to distinguish cognate alignments (shown in red) from those occurring due to the alignment of sequences from highly similar homologous loci (shown in pink).

Additional glyph details provide indication of feature properties such as transcriptional strand (forward versus reverse), clonal orientation (5' versus 3'), corresponding clone-pair sequences, annotated translational boundaries, and annotation incongruence.

From the context display, users can evaluate the level of alignment support for individual features as well as interrogate alternative features in the general vicinity. In the Figure 1 example, a researcher can ascertain that the structure of the *Zea mays* gene *TBP-2* (shown in dark blue) as defined in the GenBank record of BAC accession Z474J15 (Fig.1.6) contains an unsupported exon. This conclusion is based on the alignment of cognate cDNA and EST alignments (Fig.1 items 11 and 12). Also displayed are the alignments of homologous *Oryza sativa* protein annotations (Fig.1.10), two microarray probes (Fig.1.14), and three *Zea mays* GSS contigs (Fig.1.9) in the local vicinity of this gene annotation. A community contributed annotation (Fig.1.7, shown in green) documents one possible alternative transcript of this locus as supported by EST and cDNA alignments. A second annotation documents the downstream locus as encoding a homolog to rice gene Os3g45400, which is adjacent to the rice TBP-2 gene on rice chromosome 3, thus identifying this region as microsyntenic between maize and rice.

Genome context pages provide navigational controls allowing users to pan, zoom, and customize their view while exploring the surrounding region. Preset buttons are available to quickly zoom to a desired nucleotide resolution (Fig.1.4). The track control panel (Fig.1.5) provides a legend of the available features and controls related to their display. Display options include positional controls for altering the vertical order in which features are displayed, a visibility control for hiding the display of feature groups, filters for viewing only cognate feature alignments, and selectors for viewing extensible

glyph details such as those available with the GAEVAL extension discussed below. Adjusting the controls found in this panel will dynamically customize the genome context view without reloading the page.

Integrated web services related to the displayed genomic region are available via links (Fig.1.3) found above the context navigation controls. Typical services include display of the nucleotide sequence for the specified region, BLAST [25] query services, the yrGATE [27] community annotation tool, and a nucleotide level context page known as the transcript view. The transcript view context page displays detailed information about each feature as well as the nucleotide alignment of features derived from sequence alignment (see Figure 2). Sequences of aligned features displayed in the transcript view sequence pane use the genomic region as a scaffold to present an inferred multiple sequence alignment. Differences between feature sequences and the genomic scaffold are displayed in red to ease detection of locus defining polymorphisms and single nucleotide polymorphisms. Coordinated scrolling of the sequence alignments and the sequence view indicator allow the transcript view to provide a viewing resolution suitable to detect genome sequence base calling errors, nearby alternative splice site usage, and other nucleotide level viewing requirements without numerous page reloads

Searching and Browsing

The xGDB system provides intuitive and extensible search capabilities. Users may search for genomic locations or individual feature records using a variety of feature identifiers, aliases, keywords, or phrases entered into a common search control (Fig.1.1). Identifier searches are allowed to cascade through each feature component. Individual feature

components have the opportunity to modify the user supplied query to perform a heuristic search. For example, the official nomenclature[28] used to identify *Arabidopsis thaliana* gene annotations recommends identifiers of the form At2g42240.1. References to this gene annotation can be found at other databases under the identifiers AT2G42240.1, At2g42240 and AT2G42240. The heuristic search extensions found at AtGDB allow a user to locate this record by entering any of these identifiers.

Descriptive searches based on keywords or phrases allow users to quickly locate features of interest. A user specified search which includes phrases enclosed by quotes, keyword inclusion / exclusion operators (+ and - respectively), or which fails to locate a feature identifier will triggers a descriptive search of available feature components. Searches resulting in multiple matching features will display a summary page detailing the matching features and their genomic locations. As an example, Figure 3 shows the response to a request at AtGDB using +*"fatty acid desaturase"* -*"omega-3"*. In this query, the exclusion phrase -*"omega-3"* allows a user to narrow the results of a typical descriptive query by removing results associated with omega-3 a common class of desaturase. As described above, feature components can be individually customized to provide extended search capabilities for descriptive searches.

Evaluating feature records and their genomic alignment

Record pages provide information and web services pertinent to an individual feature. Users access record pages by clicking on a feature glyph from any context page (Fig.1.7-14) or using the record search control (Fig.1.1). Content modules, specific to each feature, control the display of record pages. These modules provide default record

displays. Providers of xGDB resources have extensive control over the customization of these modules and may even configure context page feature glyphs to link with record pages not generated by the xGDB system.

A typical record page includes information describing the feature source, peptide/nucleotide sequence(s), alignment coordinates, web service links, pertinent external website links, links to the alignment result on which the feature glyph is based, and tables summarizing the position and quality of the feature aligned to other genomic locations (see Figure 4). Display of original alignment results is a key component of xGDB which allows users to evaluate the validity of individual features as well as the method used to generate their alignment. Collection of all alignment locations and quality measures of a feature in the loci summary table allows users to quickly determine homologous genomic locations and candidate overlapping genomic sequences. Display of structure and splice site distribution glyphs for these loci provide users with interesting details on the conservation of intron size and position.

Packaged extensions

A major provision of the xGDB software design is extensibility of the core xGDB infrastructure. As such, extension of xGDB by adding third-party enhancements is encouraged. Two such enhancements, developed concurrently with xGDB, are the yrGATE gene annotation toolkit and the GAEVAL genome annotation evaluation toolkit. Both toolkits include fully functional stand-alone applications that can be incorporated into xGDB via web service extension modules.

The yrGATE toolkit provides an online portal for creation and submission of gene annotation. This web service is suitable for developing a large and nonexclusive community of annotators ranging in experience from professional curator to student. The yrGATE@xGDB extension module provides feature glyphs, search capabilities, context dependent web service links, and connections to evidence features stored in xGDB. This extension allows users to access yrGATE via web service links found on any context page for the purpose of creating an annotation. When xGDB is extended by this module additional navigational links are provided for all xGDB page headers. With these links, user can access the yrGATE annotation management pages which provide user account details, curation tools, and listings of accepted annotations.

The GAEVAL toolkit provides a system for the analysis of gene structure annotation by evaluation of supporting and incongruent evidence. This application is suitable for evaluating individual gene annotations by comparing both supporting and incongruent evidence. The GAEVAL@xGDB extension module enhances existing annotation feature components by adding glyph details to each feature cueing users as to its GAEVAL evaluation. Glyph extensions include flags for exonic sequence coverage, splice site confirmation, and possible instances of alternative splicing, alternative transcriptional termination site usage, annotation fusion, annotation fission, or erroneous annotation overlap (Fig.1.8). This web service extension also provides additional record page details (Fig.4B) about each feature evaluation as well as links to GAEVAL query and report pages.

Combining these extensions under the xGDB infrastructure establishes a framework for targeting the efforts of would-be community annotators. Through access

to the GAEVAL query service [29], lists of problematic annotations can be generated and sorted to provide a triage system for targeting annotators to interesting regions. The GAEVAL report service for each annotation can then be used to determine specific annotation alterations which are supported by current evidence. After manual evaluation of the proposed alterations, an annotator may use the yrGATE service [30] to provide an updated gene structure annotation. Upon acceptance of this user contributed annotation, the GAEVAL system is used to re-evaluate the current annotation thereby documenting the presence of the new yrGATE submission.

xGDB INTERNALS

We now describe the internal design and back-end components of xGDB accessible to data providers and users desiring private installations. We first present the overall system design which has focused on modularity and extensibility. Options for integrating alternative database structures and distributed database architecture are discussed next. We then detail the feature component modules that are distributed with xGDB. And finally, we discuss options for installation and custom configuration of an xGDB system.

Software design, modularity, and extensibility

The xGDB system consists of both user interface and data management components. Together these components make xGDB highly modular and extensible. On the front-end, the xGDB user interface is provided by a collection of CGI (Common Gateway Interface) scripts. Core CGI scripts are maintained in data independent modules such that multiple xGDB systems may be operated using a single core installation. The AtGDB and

ZmGDB systems illustrated herein as well as all other species configurations maintained by PlantGDB operate from a single xGDB core by taking advantage of this design feature. In addition, extended functionality such as that of the GAEVAL@xGDB service can be installed in a centralized location and made optionally accessible to all local xGDB systems.

Data management and back-end database interoperability are provided by the xGDB database object and independent feature component modules. Modular feature components allow plug-in like inclusion of new feature sources as well as customization of existing sources. Feature components are built from an object oriented paradigm where required methods are gained through object inheritance and can be customized or extended by overriding individual method instances. These overrides may take place in either the component class or individual instances of an existing class. Figure 5 depicts the object structure and point of customization of two features in use at AtGDB. The GenBank mRNA annotation feature uses a standard GenBank feature component which has been customized by addition of GAEVAL specific method instances. For this component, the underling class itself was altered. The PlantGDB Unique Transcript feature however uses a standard cDNA feature component and is customized simply by addition of a modification file. This design allows for expansion and a variety of features to be uniquely represented with minimal additional effort.

Integration with distributed and federated database systems

The xGDB database object manages the individual component features and provides adaptor methods for the relational database system of each component. Using an adaptor

methodology, the choice of database management system, host, and scheme can be delegated to each feature component. As such xGDB is capable of operating under distributed database architectures. One highly appealing use for such architecture is in maintaining an often changing feature set. For instance, local use of the individual EST and cDNA alignment feature available at AtGDB would necessitate a pipeline for continuous update as new sequences become available. This poses a challenge both in resource and time commitment for most small to moderately sized research groups. The ability of xGDB to utilize a distributed architecture however allows PlantGDB to provide direct connection to available PlantGDB feature sources (see Table 1). Therefore an individual xGDB maintainer need only configure their xGDB system to utilize this connection in order to remain up-to-date with the features found at PlantGDB.

The variety of genomic features, distribution sources, and distributed formats currently available for genomic context analysis necessitates an infrastructure system with federated data management capabilities. The modular design of xGDB allows creation of feature components specific to any distribution source or format. In addition to its native database architecture, the xGDB system is currently capable of using DAS (Distributed Annotation System) [31] distribution sources and GFF (General Feature Format) databases [8] by providing feature component modules with federated data management adaptors. This allows integration with widely available tools and data distributed by projects such as Ensemble and GMOD. Examples and instructions for using these adaptors are provided with the xGDB installation notes.

Feature component modules

Feature component modules consist of a Perl encoded DSO (data source object), web service scripts providing unique functionality to each feature component, data management scripts for loading features from flat files of various formats into a relational database management system, and supporting information necessary for feature configuration and customization. A variety of modules are available in the core xGDB distribution including those encapsulating GenBank gene features, TIGR transcription units, and GeneSeqer expressed sequence spliced alignments. Incidentally, any genomic feature which can be positioned by a genomic coordinate can be developed into a feature component module. For example, with only minor modification of existing modules we have added predicted repeats, GSS (genome survey sequence) alignments, and microarray probe positions to the feature component modules in use at PlantGDB. As described in the following, existing feature component modules and their common DSO design provides an ample infrastructure for managing most genomic features.

The DSO of each modular feature component inherits from a rich object framework which allows efficient method inheritance and less coding to develop objects encompassing new genomic feature sources (see Figure 5). Currently all DSOs descend from the Locus base object which instantiates required object methods and provides a common object constructor. Most DSOs inherit the Locus object through hierarchical inheritance from second-tier objects such as the Annotation, Sequence, DAS, or BioDBGFF objects. These objects contribute standardized routines for searching, display, and interaction with feature components derived from each respective category. DSO are often enhanced through multiple inheritance as is the case with the cDNA and EST

objects shown in Figure 5 which inherit both from the Sequence object and the GeneSeqerSequence object.

Method callbacks and subroutine hooks are used in the DSO framework to allow single instance customization of often modified object methods such as identifier and descriptive search routines, context region and record link publishers, and feature information HTML generators. The methods inherited from either the Annotation or Sequence objects encode subroutine hooks which allow a DSO to be customized by declaring a 'mod' file as an object configuration parameter. When declared, this 'mod' file is included in the DSO framework for its respective feature component. In Figure 5, the GAEVAL enhanced GenBank gene feature DSO is shown to use a 'mod' file which provides an identifier validation routine responsible for heuristically altering a user supplied query to match feature identifier formats as found in the underlying MySQL database. The PUT (PlantGDB Unique Transcript) DSO also uses a 'mod' file. This modification however, is used to alter the cDNA DSO instance thereby allowing it to encapsulate the PUT feature component.

Installing and customizing xGDB

Setting up an xGDB system requires installation of the core xGDB distribution, installing an xGDB instance, populating a feature component module, and configuring the xGDB instance to include the feature component. Documentation and installation scripts are provided with the xGDB distribution to expedite this process. Instances are generally populated with multiple feature components. Components are associated with each xGDB instance through an instance configuration file. Additional xGDB instances can be

configured for additional species or separation of publicly accessible resources from proprietary systems. Each subsequent instance may share the initial xGDB core and any feature components installed therein. Instance based customization of feature component modules as described above may be used to further distinguish individual xGDB resources.

Extensive options for customizing an xGDB instance are available. User interface properties such as color, image logos, and page layout are determined using a cascading style sheet. Modification of the default style sheet provided in the xGDB distribution allows an xGDB installer to quickly give any instance a unique look. Site navigation menus and controls can be customized using instance configuration files as well. These customization options are used with the xGDB instances at PlantGDB to provide additional informative content. This content includes species specific download pages, web pages relating relevant projects involving the use of xGDB such as the characterization of U12-dependent introns using AtGDB, and links to relevant websites maintained by other research organizations.

The xGDB distribution is available for download [32] and requires only widely-available open-source software. All distributed modules and required software run well on a variety of Unix-based systems including Linux and Macintosh OS X. The xGDB systems interact with end-users through a combination of PHP and PERL generated web pages. Internet browsers that support HTML level 4, core JavaScript version 1.4 and higher, and Cascading Style Sheets level 2 and higher are required for complete user interface functionality. Default web pages have been design tested using Mozilla Firefox version 1.5.

CONCLUSIONS

The xGDB system provides an infrastructure for organization of genomic data, analysis of a wide range of inquiries about such data, and online publishing of both the data and analysis results. The extensible design of xGDB provides a packaged solution to many types of research applications. In particular, xGDB is well suited for small to moderately sized research groups desiring local access to genomic data or an out-of-the-box system for analyzing emerging data.

xGDB SOFTWARE REQUIREMENTS

The xGDB system requires the following software packages:

1. The Apache Web server [33], version 1.3 or higher
2. The PHP apache server API [34], version 3 or higher
3. The Perl interpreter [35], version 5 or higher
4. The following Perl modules found at CPAN [36]: DBI, DBD::mysql, GD, CGI

xGDB SUPPORT

The xGDB project is hosted on SourceForge.net, an online open-source development community. The complete xGDB distribution can be obtained from the xGDB project website [37]. This site includes utilities for user support, versioned distribution releases, bug reports, and feature requests. Forums at this site are regularly monitored by xGDB developers. The PlantGDB site also provides a user feedback utility to assist in user

support for PlantGDB resources and requests. Links to this utility can be found in the header of all PlantGDB maintained web pages.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Plant Genome Research Program grant DBI-0321600 to V.B. Shannon Schlueter was supported in part by the National Science Foundation Integrative Graduate Education and Research Traineeship (IGERT) grant DGE-9972653.

REFERENCES

1. Butler D, Smaglik P: **Draft data leave geneticists with a mountain still to climb.** *Nature* 2000, **405:**984-985.
2. Stein LD: **Using Perl to facilitate biological analysis.** *Methods Biochem Anal* 2001, **43:**413-449.
3. Field D, Feil EJ, Wilson GA: **Databases and software for the comparison of prokaryotic genomes.** *Microbiology* 2005, **151:**2125-2132.
4. Rajpal DK: **Understanding biology through bioinformatics.** *Int J Toxicol* 2005, **24:**147-152.
5. Stein L: **Genome annotation: from sequence to biology.** *Nat Rev Genet* 2001, **2:**493-503.
6. Howe KL, Chothia T, Durbin R: **GAZE: a generic framework for the integration of gene-prediction data by dynamic programming.** *Genome Res* 2002, **12:**1418-1427.

7. Gilbert D: **Bioinformatics software resources.** *Brief Bioinform* 2004, **5:**300-304.

8. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12:**1599-1610.

9. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, et al: **Ensembl 2006.** *Nucleic Acids Res* 2006, **34:**D556-561.

10. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2006, **34:**D16-20.

11. Usuka J, Zhu W, Brendel V: **Optimal spliced alignment of homologous cDNA to a genomic DNA template.** *Bioinformatics* 2000, **16:**203-211.

12. Dong Q, Lawrence CJ, Schlueter SD, Wilkerson MD, Kurtz S, Lushbough C, Brendel V: **Comparative plant genomics resources at PlantGDB.** *Plant Physiol* 2005, **139:**610-618.

13. Zhu W, Schlueter SD, Brendel V: **Refined annotation of the Arabidopsis genome by complete expressed sequence tag mapping.** *Plant Physiol* 2003, **132:**469-484.

14. Zhu W, Brendel V: **Identification, characterization and molecular phylogeny of U12-dependent introns in the Arabidopsis thaliana genome.** *Nucleic Acids Res* 2003, **31:**4561-4572.

15. Schlueter SD, Wilkerson MD, Huala E, Rhee SY, Brendel V: **Community-based gene structure annotation.** *Trends Plant Sci* 2005, **10:**9-14.

16. Wang BB, Brendel V: **Genomewide comparative analysis of alternative splicing in plants.** *Proc Natl Acad Sci U S A* 2006, **103:**7175-7180.

17. Wang BB, Brendel V: **Molecular characterization and phylogeny of U2AF35 homologs in plants.** *Plant Physiol* 2006, **140:**624-636.

18. Ashurst JL, Collins JE: **Gene annotation: prediction and testing.** *Annu Rev Genomics Hum Genet* 2003, **4:**69-88.

19. Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK, Jr., Maiti R, Chan AP, Yu C, Farzad M, Wu D, et al: **Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release.** *BMC Biol* 2005, **3:**7.

20. Yuan Q, Ouyang S, Wang A, Zhu W, Maiti R, Lin H, Hamilton J, Haas B, Sultana R, Cheung F, et al: **The institute for genomic research Osa1 rice genome annotation database.** *Plant Physiol* 2005, **138:**18-26.

21. Hong P, Wong WH: **GeneNotes--a novel information management software for biologists.** *BMC Bioinformatics* 2005, **6:**20.

22. **AtGDB** [http://www.plantgdb.org/AtGDB/]

23. **ZmGDB** [http://www.plantgdb.org/ZmGDB/]

24. **PlantGDB** [http://www.plantgdb.org/]

25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

26. Schlueter SD, Dong Q, Brendel V: **GeneSeqer@PlantGDB: Gene structure prediction in plant genomes.** *Nucleic Acids Res* 2003, **31:**3597-3600.

27. Wilkerson MD, Schlueter SD, Brendel V: **yrGATE: a web-based gene-structure annotation tool for the identification and dissemination of eukaryotic genes.** *Genome Biology* 2006, **in press**.

28. **TAIR Nomenclature Guidelines** [http://www.arabidopsis.org/portals/nomenclature/guidelines.jsp]

29. **GAEVAL @ AtGDB** [http://www.plantgdb.org/AtGDB-chtml/GAEVAL.php]

30. **yrGATE @ AtGDB** [http://www.plantgdb.org/AtGDB_yrGATE-cgi/CommunityCentral.pl]

31. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L: **The distributed annotation system.** *BMC Bioinformatics* 2001, **2:**7.

32. **xGDB** [http://xgdb.sourceforge.net/]

33. **Apache Web Server** [http://www.apache.org/]

34. **PHP** [http://www.php.net/]

35. **PERL** [http://www.perl.org/]

36. **CPAN** [http://www.cpan.org/]

37. **xGDB project at Sourceforge** [http://sourceforge.net/projects/xgdb/]

38. **MySQL Boolean Full-text Searches** [http://dev.mysql.com/doc/refman/5.0/en/fulltext-boolean.html]

Table 1. Feature sources provided by PlantGDB.

Species	Genomic Sequences			Annotations		Expressed Sequences			
	Chr	BAC	GSS	GenBank	yrGATE	EST	cDNA	PUT	Probe
A. thaliana	5	-	-	34513	29	622788	66445	144274	251078
B. rapa	-	52	-	-	-	21222	381	13040	-
G. max	-	66	-	-	-	358702	1116	101998	671762
L. esculentum	-	89	-	467	-	199873	3291	40966	112528
L. japonicus	-	1374	-	170	-	149878	224	43592	-
M. truncatula	-	1644	-	18971	-	225129	787	54395	673880
O. sativa	12	3462	-	68761	6	406790	35318	141239	631066
P. trichocarpa	-	173	-	-	-	89943	119	29640	-
S. bicolor	-	41	79343	-	-	204208	110	44958	-
T. aestvum	-	57	-	-	-	853621	2386	243326	-
Z. mays	-	2031	294425	936	10	714484	14476	140616	57452

Chr: Chromosome; BAC: Bacterial Artificial Chromosome; GSS: Genome Survey Sequence; EST: Expressed Sequence Tag; cDNA: complementary DNA; PUT: PlantGDB Unique Transcript; Column values represent the number of unique features/sequences made available at PlantGDB. The protein column represents the sum of all cross-species homologous protein alignments. Each expressed sequence may be responsible for multiple features by alignment to multiple loci.

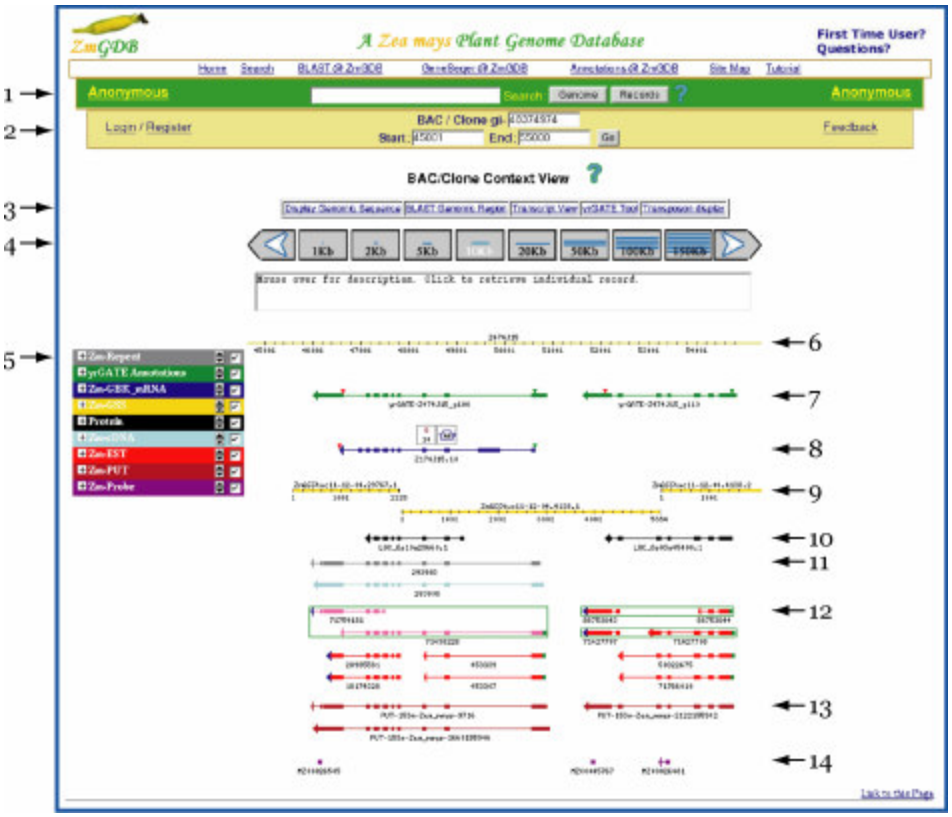

Figure 1. A ZmGDB context page focused on a *Zea mays* BAC assembly (accession Z474J15, GenBank id 48374974). A site header contains site navigation and search controls (1&2). Links to integrated webservices (3) and context navigation controls (4) are available. The feature control panel (5) and context graphic shows yrGATE community annotations (7), GenBank gene features (8), PlantGDB GSS assemblies (9), rice predicted protein alignments (10), cDNA alignments (11), EST alignments (12), PlantGDB Unique Transcript alignments (13), and MaizeArray microarray probe alignments (14) in the genomic region spanning bases 45001 to 55000 (6) of the assembled sequence. Exon features are displayed as filled rectangles connected by intronic features represented by similarly colored lines. Predicted start and stop codons of open reading frames are represented by green and red triangles respectively. Arrowheads represent genomic strand orientation when this can be determined. Noncognate features are represented by alternative feature colors (pink for EST and grey for cDNA features).

Figure 2. A ZmGDB transcript view context page associated with the genomic region depicted in Figure 1. The feature graphic in the top window pane is described in Figure 1. Information at the top and left of this pane is displayed when passing the cursor over feature elements. Currently displayed is the information associated with the sixth intron (immediately left of the green viewfinder) of the GeneSeqer spliced alignment of a *Zea mays* cDNA sequence (accession AV109414, GenBank id 21213129). The vertical green bars represent the view finder for the sequence view found in the bottom window pane. Red nucleotides shown in this view represent alignment mismatches with the genomic sequence.

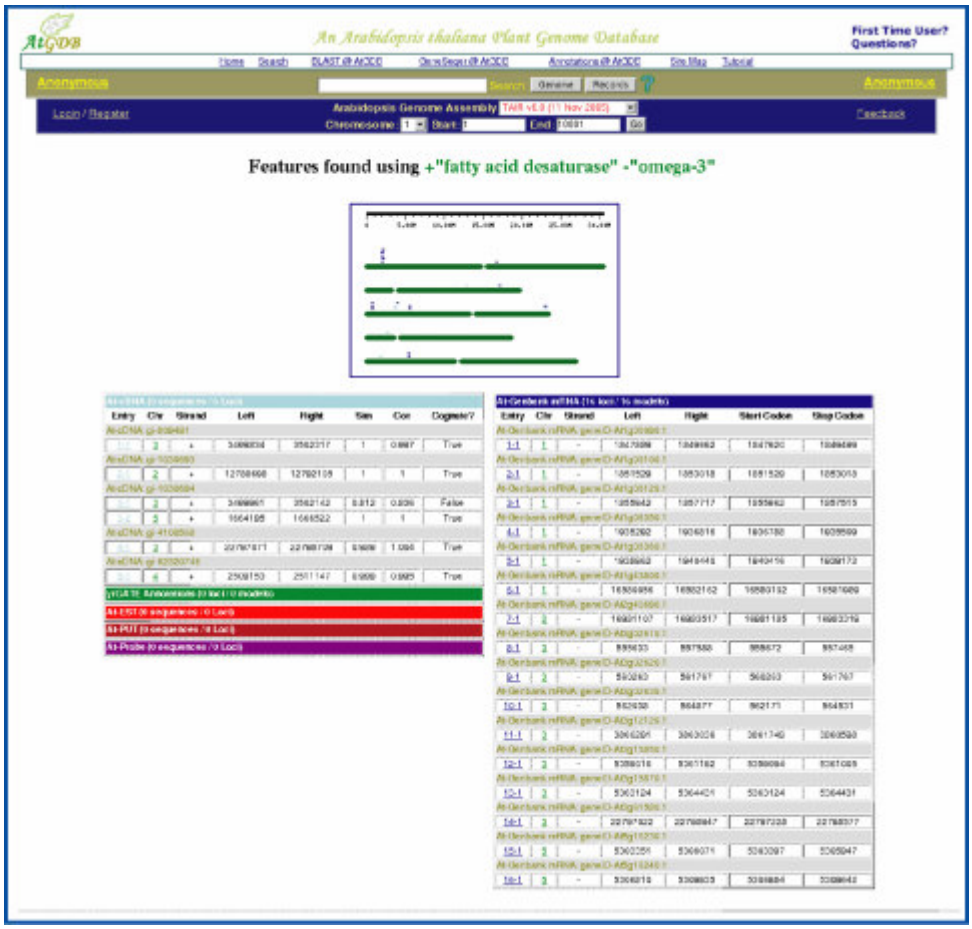

Figure 3. Search results at AtGDB using the query *+"fatty acid desaturase" -"omega-3"*. The "+" and "-" operators represent inclusion and exclusion, respectively, following the convention of MySQL boolean text searches [38].

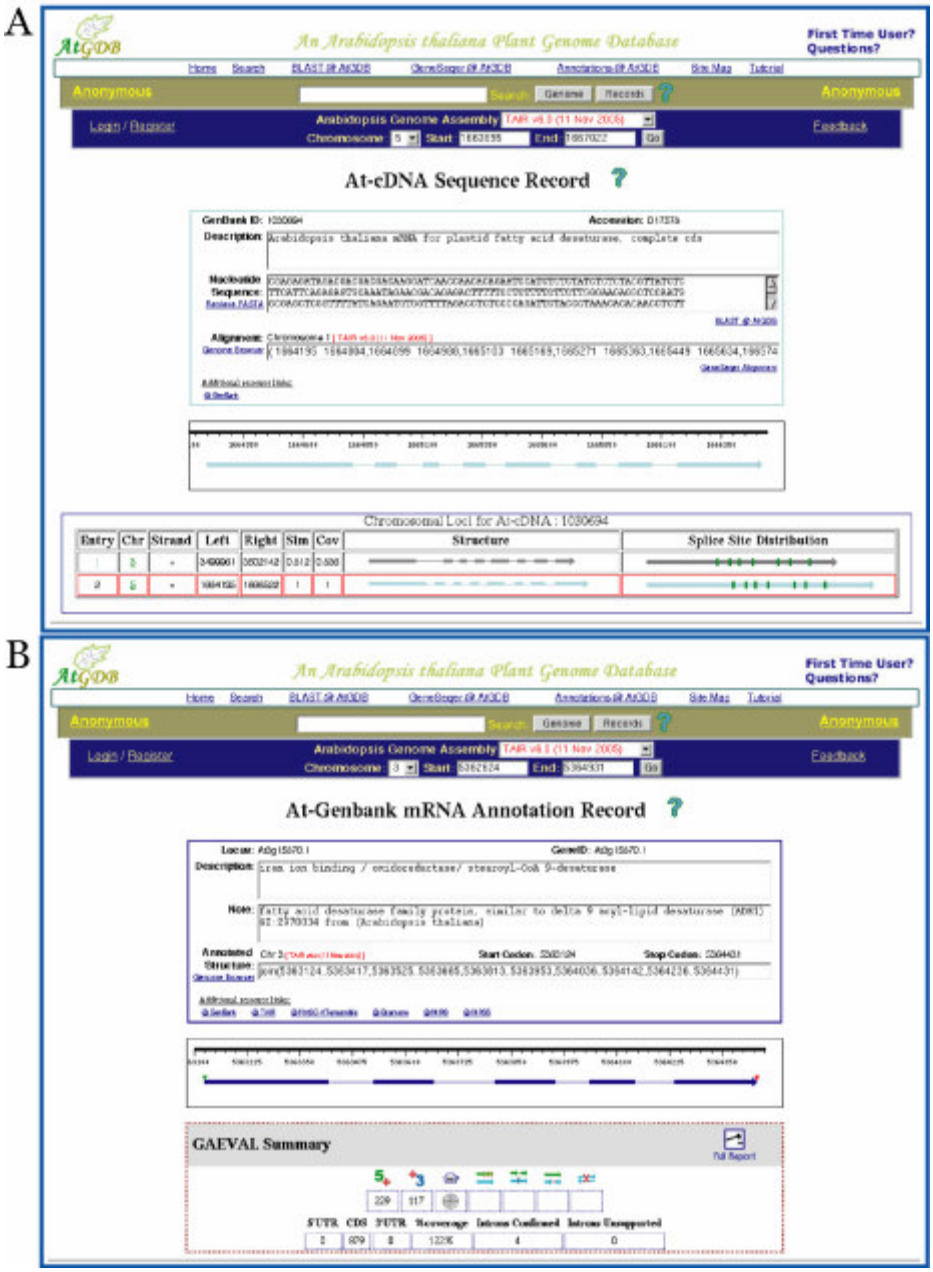

Figure 4. [A] An AtGDB record page summarizing the GeneSeqer spliced alignment of an *Arabidopsis thaliana* cDNA sequence (accession BT020201, GenBank id 55733740). Feature structure glyphs found in the alignment loci summary table at the bottom of the window are as described in Figure 1. Green bars in the splice site distribution glyph represent the location of slice junctions in the processed messenger RNA transcript. **[B]** An AtGDB annotation record page detailing an *Arabidopsis* gene annotation (At3g15870.1). The GAEVAL Summary report at the bottom of the window displays information obtained using the integrated GAEVALxGDB services.

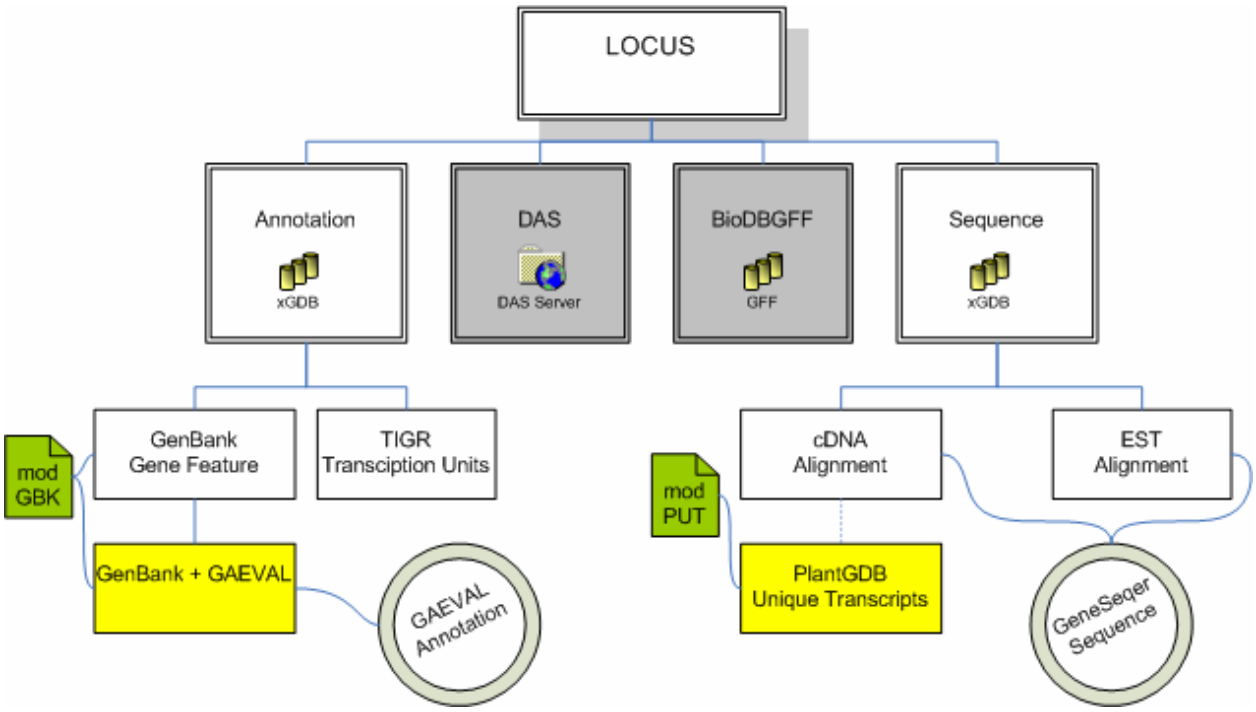

Figure 5. A partial representation of the object model for data source objects (DSO) being used at AtGDB. Customized features derived from distribution objects are shown in yellow. Solid lines represent object inheritance. The dashed line connecting the PlantGDB Unique Transcripts feature represents instantiation of the cDNA DSO. Grey objects represent federated adaptors to resources developed elsewhere.

CHAPTER 3: COMMUNITY-BASED GENE STRUCTURE ANNOTATION

A paper published in *Trends in Plant Science*[1]

Shannon D. Schlueter[2], Matthew D. Wilkerson[2], Eva Huala[3], Seung Y. Rhee[3] and Volker Brendel[2,4]

Abstract

Uncertainty and inconsistency of gene structure annotation remains a limitation to research in the genome era, frustrating both biologists and bioinformaticians who are forced to spend considerable and often duplicated efforts to sort out annotation errors for their genes of interest or to generate trustworthy data sets for algorithmic development. It is unrealistic to hope for better software solutions in the near future that would solve all the problems. The issue is all the more urgent with more species being sequenced and analyzed by comparative genomics – erroneous annotations could easily propagate, whereas correct annotations in one species will greatly facilitate annotation of novel genomes. We propose a dynamic and economically feasible solution to the annotation predicament: broad-based, web technology enabled community annotation. Such a system has now been implemented for Arabidopsis and is easily portable for use in other species-specific resources.

---

[1] Reprinted with permission of *Trends in Plant Science,* 2005, 10, 9-14.
[2] Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA
[3] Carnegie Institution, Department of Plant Biology, Standford, CA 94305, USA
[4] Department of Statistics, Iowa State University, Ames, IA 50011, USA, corresponding author

When is a genome finished?

For all plant and animal species, presentation of the "finished genome" is considered a major milestone in the study of its genetics. Ambiguous claims of this highly prized accomplishment, however, beg the question as to the meaning and worth of such announcements. Competitive and controversial claims concerning the completion of the human genome have been widely discussed [1]. In the area of plant genetics, the completed *Arabidopsis thaliana* genome was reported at the end of 2000 [2]. At that time, the genomic assembly comprised 115,409,949 base pairs covering the five chromosomes and leaving only an estimated 10 megabases of centromeric and rDNA repeat regions not sequenced. The total length of the assembled genome has increased by about one megabase per year (http://www.plantgdb.org/AtGDB/resources.php). A more demanding definition of a "finished genome" requires extensive annotation of the assembled chromosome sequences in addition to the mere sequence report. In particular, researchers using the genome as a model system require annotation of the protein coding genes as the basis for assessing the transcriptome and proteome of the species. At the time of the Arabidopsis genome release, 25,498 protein-coding genes were annotated on the genome sequence. Since that time, this annotation challenge has continued to receive serious consideration for *Arabidopsis*, as evidenced by an approximate 10% increase in the number of annotated gene structures during the last three years [3] and continuing correction of erroneous initial annotations [4]. Perhaps the most ambitious and accurate definition of a "finished genome" should include functional characterization of all the genes, a goal of the Arabidopsis 2010 project [5]. It is clear that each successively more

comprehensive definition requires completion of the less ambitious tasks. The complexities of providing comprehensive annotation, whether that annotation is structural or functional, depend on an accurately defined gene structure. Because our collective understanding of genes and genome function continually advances, and users of the genome annotation naturally expect it to remain up-to-date with recent discoveries, the definition of a finished genome is, of necessity, something of a moving target.

Currently, a considerable time lag between completion of sequencing and completion of annotation appears unavoidable. This is because even though sequencing is largely automated and robotic and sequence assembly is largely routine (at least for genome regions that are not highly repetitive), accurate sequence annotation entirely by gene finding software has remained elusive [6]. Current efforts toward more accurate and comprehensive gene structure annotation have focused on EST and full-length cDNA mapping onto the *Arabidopsis* genome [7-9] and combinations of computational and experimental approaches [10-11]. These studies have underscored the utility of spliced alignment to identify non-coding exons and to correct inaccurate computational gene predictions that formed the basis of the initial genome annotation. In particular, the results of cDNA mapping point to inherent limitations of high-throughput computational gene prediction, including difficulties in predicting exact exon borders, problems with distinguishing intergenic regions from introns, and lack of models capable of identifying untranslated mRNA regions. However, these recent efforts have also not been entirely immune to the problems of large-scale automated annotation. For example, novel algorithmic changes incorporated into the newest annotation release [12] have

inadvertently resulted in the ambiguous assignment of Expressed Sequence Tags (ESTs) to multiple adjacent genes, thereby falsely extending their gene structure annotations [e.g., Fig. 2 in ref. 13]. Inclusion of draft sequences of clones too repetitive to finish with existing technology, while useful as a way to improve genome coverage with the available fragments of sequence data, has had some undesirable consequences, such as inclusion of pBlueScript vector sequences in the genome sequence (http://www.plantgdb.org/AtGDB/Annotation/vector.php). The scope and complexity of the genome annotation task would seem to imply that shortcomings and mistakes are simply unavoidable in the early to middle stages of finishing a genome. Hild et al. [14] discuss similar challenges with respect to the *Drosophila* genome annotation.

Arabidopsis genome annotation

The Arabidopsis research community currently has several ways to access genome data. TAIR [The Arabidopsis Information Resource; http://arabidopsis.org; ref. 15], TIGR [The Institute for Genome resources; http://www.tigr.org/tdb/], MATdb [MIPS Arabidopsis thaliana Databases; http://mips.gsf.de/proj/thal/db/; ref 16], SIGnAL [Salk Institute Genomic Analysis Laboratory; http://signal.salk.edu/; ref 17], and AtGDB [The *Arabidopsis thaliana* genome database at PlantGDB; http://www.plantgdb.org/AtGDB; ref. 9, 13] provide web-based genome browsers for *Arabidopsis* that display gene structure annotation and comparisons with spliced alignment of ESTs and cDNAs. In addition to its genome browser, TAIR provides a comprehensive access point for *Arabidopsis* data, including information about genes, sequences, proteins, microarrays, germplasms, polymorphisms, seed and DNA stocks, and the research community.

TAIR's curation efforts include the functional annotation of genes, with an emphasis on capturing experimental data from the literature and using controlled vocabularies [18].

Since the first release of the genome sequence in 2000, the Institute for Genome Resources (TIGR) has maintained and updated the Arabidopsis genome annotation, making the updates publicly available in periodic releases ending with the TIGR 5.0 release in January 2004, visible at both AtGDB and TAIR. As TIGR's role in maintaining and improving the genome annotation has come to an end, other mechanisms must be put in place to insure that the genome data remain as error-free and up to date as possible. In response to this need, TAIR is currently setting up its own automated pipeline for improving gene models using new EST and cDNA data and manual methods for updating gene structures in response to community input. While TAIR will work to eliminate the previously reported problems associated with automated gene structural annotation, automated methods will never be as flexible as a human curator in handling unusual cases or making use of new kinds of data. However, manual curation efforts by trained curators are limited by the size of the curation team and the amount of time needed to resolve each problematic gene structure annotation.

Even with well-organized community resources to support the informatics needs of a genome project, genome annotation remains a difficult task because ultimately all gene models will have to be evaluated by human experts. We have argued previously [19, 20] that the only promising solution to this quandary is involvement of the user community and development of enabling technology that streamlines user input, curation of user

contributions, and dissemination of approved user contributions. The purpose of this article is to introduce web-based gene structure annotation tools that are directly linked into AtGDB and TAIR and will, we believe, facilitate broad-based community participation in the genome annotation task.

To assist in evaluating the quality of specific gene structure annotation and to determine the overall quality of the current *Arabidopsis* annotation, we have developed a system at AtGDB that allows gene structure comparison in the genomic context (http://www.plantgdb.org/AtGDB/Annotation/). The system, named Genome Annotation EVALuation (GAEVAL, pronounced **gavel** [**g**$\breve{a}$**v**'$\breve{e}$**l**]), highlights inconsistencies between current gene structure annotation and the cognate placement of spliced aligned ESTs and (full-length) cDNAs. The reference for current gene structure annotation is provided by the mRNA fields in the GenBank deposited chromosome sequence files (accessions NC_003070, NC_003071, NC_003074, NC_003075, NC_003076). The cognate spliced alignments were derived with the GeneSeqer program as described previously [9] and provide the utility to identify non-coding exons, confirm splicing boundaries, and correct inaccurate *ab initio* gene predictions. Additionally, due to the nature of cognate mapping, these spliced alignments provide higher accuracy when evaluating genes from multigene families by explicitly utilizing only sequences native to the specific locus for annotation.

Quality assessment of predicted gene structures

Alignments are first evaluated to determine their native locus and, if necessary, the specific transcript isoform derived from the locus. A scoring system comparing the

spliced alignment with overlapping gene annotations was devised to aid in this determination (http://www.plantgdb.org/AtGDB/Annotation/gaeval/). Once a transcript isoform has been identified from which the EST or cDNA originated, all corresponding spliced alignments are compared to the predicted gene structure. This comparison is used to judge the accuracy of the gene annotation and to assign a quality flag for immediate appraisal of annotation validity. Five levels of annotation quality were established as shown in Fig. 1. The first quality level corresponds to an unconfirmed gene annotation for which currently no EST or cDNA evidence is available. These gene structure annotations are generally based entirely on *ab initio* computational prediction. Further analysis using homologous ESTs and cDNAs can be used to provide estimates of the annotation accuracy [21, 22]. Annotations of quality levels beyond the first benefit from the spliced alignment of ESTs and cDNAs. Increasing quality levels, as described in Fig. 1, are representative of increasing confidence in the accuracy and completeness of an annotation. Ultimately, the fifth level quality assignment is given to gene annotations completely tiled by cognate ESTs or cDNAs, with all splice site boundaries supported. These annotations represent well supported gene structures that have the least anticipated need of future modification.

Inconsistencies found by this comparison are used to flag possible alternative splicing and gene structure deviations as well as inaccurate prediction of introns and intergenic regions (Fig. 2, http://www.plantgdb.org/AtGDB/Annotation/gaeval/gaeval_lists.php). Alternative splicing is made evident when the supported gene structure of a given locus is incongruent with that of the spliced alignment of one or more native expressed

sequences. Validation of an alternative isoform can be based on criteria such as the number of ESTs and cDNAs supportive of the alternative structure, the acceptability of the alternative splice junction relative to known models, and the surrounding context of the alternative isoform (i.e., proper open-reading frame, presence of splicing enhancers, etc.). Consistent alignments can provide clues as to incomplete or inaccurate annotation as well. For example, an expressed sequence alignment also matching to adjacent non-overlapping gene annotations is common evidence of a falsely predicted gene termination or intergenic region (e.g., center example in Fig. 2). This mistake creates separate gene annotations representing fragments of a single gene. In addition, consideration of sequence vector properties, such as the source clone of an EST or the 5' versus 3' origination of the EST from the clone, can aid in determining the extent of a valid gene structure. Clone-pair ESTs (ESTs obtained from opposite ends of a cDNA clone) provide an often-overlooked indicator of fragmented gene structure annotation (Fig. 3). Another less common mistake, whereby independent gene structures are incorrectly fused into a single gene annotation, can be caught using clone-pair ESTs, groups of 3' ESTs, and "full-length" cDNAs as evidence (http://www.plantgdb.org/AtGDB/Annotation/gaeval/cps.php). Though they require extremely robust algorithms to correct in an automated fashion, these anomalies are easily found and corrected by manual curation when presented appropriately.

Genomic context visualization, as provided at AtGDB [9, 13], can be used to correct annotation mistakes for which automated correction is not feasible and to validate the behavior of novel automated annotation routines. For example, the inaccurate extension

of some *Arabidopsis* gene models incorporated in the latest annotation release apparently resulted from changes in the automated annotation routines now in use [12]. Context visualization makes it possible to find these anomalies in a targeted and user accessible manner (Fig. 3). In addition, the genome context visualization and display of user comments allows for more complex annotation than can be captured with current GenBank feature tags. For example, some proportion of the gene structure pairs flagged as potentially needing to be merged by cDNA evidence correctly represent distinct translation products derived from dicistronic mRNAs (C. Town, personal communication). Standards for annotating such cases have not been set, and therefore these cases are currently not represented in GenBank feature tags. It is also clear that such complex cases would be very difficult to annotate automatically. The success of automated annotation pipelines relies on stringent criteria that capture the most reliable annotations [12]. In our view, the subsequent phase of completing the annotation can only be achieved by broad-based community input.

Community-based annotation

To insure maximum participation by the community, the tools for updating annotations must be accessible and convenient for those viewing the data. Since TAIR is a heavily used resource, many users will notice a structural annotation problem first in TAIR. In addition, TAIR users already routinely submit comments and corrections on a variety of types of data using a comment field on TAIR detail pages or email directly to TAIR curators, suggesting that the TAIR user community is willing to contribute information. Therefore, we have added a link on TAIR data pages where gene structural annotation

information is visible, allowing users wishing to correct a gene structure to automatically connect to AtGDB's GAEVAL system. By connecting TAIR and AtGDB through use of a centralized authentication service, we can enable a TAIR user's identity to be securely passed to AtGDB ensuring proper attribution. The corrected structures are sent back to TAIR automatically where they will be checked by a curator before being incorporated into the next version of the genome. The TAIR curator will examine the updated gene using the Apollo genome annotation tool [23] to confirm that existing cognate cDNAs and ESTs support the new gene model, verify the translational start and stop for protein coding genes, and review and update any functional annotation attached to the gene. This review will insure that the new annotation conforms to TAIR's curation standards. If TAIR curators detect a consistent pattern of error in user-submitted annotations, AtGDB will make use of this information to improve the interface to prevent the error, either by improving the tools available to the submitter or by alerting the submitter of the error at the time of submission.

Outreach

We believe that genome annotation could be an excellent vehicle for education, both at the high school and undergraduate levels. To this end, we have developed a tutorial site at AtGDB (http://www.plantgdb.org/AtGDB/tutorial/) that guides users through the terminology and practice of gene structure annotation. This development was achieved in collaboration with local high schools in the Iowa State University area. In addition, talented high-school student interns have proven to be both eager and effective users of these gene annotation tools and have greatly contributed to improvements in tool design

and scope (see http://www.plantgdb.org/AtGDB/Interns/ for project descriptions and results).

Conclusions

The community curation approach has the potential to solve the problem of how to maintain a high quality genome annotation for the long term. While some of the problems with existing automated annotation pipelines might eventually be corrected, manual curation remains the best method for producing high quality genome annotation. The tools and resources presented here have made such community curation convenient and efficient while providing wide access to the resulting data. Greater than 200 such community annotations are currently accessible at AtGDB (see http://www.plantgdb.org/AtGDB/Annotation/UCAlist.php).

Acknowledgements

This work was supported by NSF Plant Genome Research Grant DBI-0321600 to V.B. and NSF grant number DBI-9978564 to S.Y.R. We would like to thank Mr. Michael Lawler and Ms. Stephanie Haila, both science school teachers in Iowa, for working with us to develop the gene structure annotation tutorial. Their work was sponsored by an NSF RET grant to ISU. We would also like to thank Dr. Chris Town of The Institute for Genomic Research (TIGR) for critical reading and comments.

References

1. Roberts, L. (2001) Controversial From the Start. *Science* 291, 1182-1188.
2. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796-815
3. Wortman, J.R. *et al*. (2003) Annotation of the Arabidopsis genome. *Plant Physiol.* 132, 461-468
4. Brendel, V. and Zhu, W. (2002) Computational modeling of gene structure in *Arabidopsis thaliana*. *Plant Mol. Biol.* 48, 49-58
5. Ausubel, F.M. (2002) Summaries of National Science Foundation-sponsored Arabidopsis 2010 projects and National Science Foundation-sponsored plant genome projects that are generating Arabidopsis resources for the community. *Plant Physiol.* 129, 394-437
6. Pavy, N. *et al*. (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics* 15, 887-899
7. Seki, M. *et al*. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* 296, 141-145
8. Haas, B.J. *et al*. (2002) Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biology* 3, research0029.1-0029.12
9. Zhu, W. *et al*. (2003) Refined annotation of the *Arabidopsis thaliana* genome by complete EST mapping. *Plant Physiol.* 132, 469-484
10. Yamada, K. *et al*. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* 302, 842-846

11. Castelli, V. *et al*. (2004) Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation. *Genome Res*. 14, 406-413

12. Haas, B.J. *et al*. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654-5666

13. Dong, Q., *et al*. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.* 32, D354-D359

14. Hild, M. *et al*. (2003) An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol*. 5, R3.1-R3.16

15. Rhee, S.Y. *et al*. (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res*. 31, 224-228

16. Schoof, H. *et al.* (2002) MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res.* 32, D373-D376

17. Yamasa, K. *et al.* (2003) Empirical Analysis of Transcriptional Activity in the Arabidopsis Genome. *Science* 302, 842-846.

18. Berardini, T.Z. *et al*. (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol.* 135, 745-755

19. Rhee, S.Y. (2004) *Carpe Diem*: Retooling the 'Publish or Perish' Model into the 'Share and Survive' Model. *Plant Physiol.* 134, 543-547

20. Brendel, V. (2003) Novel tools for plant genome annotation and applications to *Arabidopsis* and rice. In J.P. Gustafson, R. Shoemaker and J.W. Snape (eds.), "Genome Exploitation: Data Mining", *Stadler Genetics Symposia Series, 23rd Symposium*, Kluwer Academic/Plenum Publishers, U.S.A., to appear.

21. Schlueter, S.D. *et al*. (2003) GeneSeqer@PlantGDB - gene structure prediction in plant genomes. *Nucleic. Acids Res.* 31, 3597-3600

22. Brendel, V. *et al*. (2004) Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics* 20, 1157-1169

23. Lewis, SE. *et al.* (2002) Apollo: a sequence annotation editor. *Genome Biol.* 3, research0082.1-research0082.14

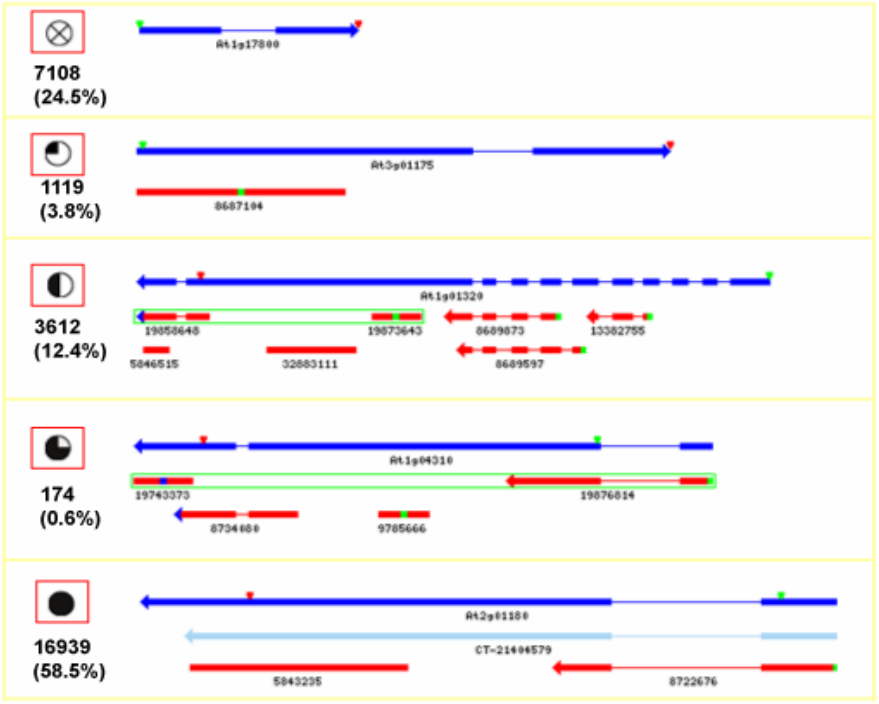

Figure 1. Levels of support for gene structure annotation. This graphic uses existing gene annotations and their depiction (as found at http://www.plantgdb.org/AtGDB/) to illustrate varying levels of confirmation evidenced by spliced alignment of cognate cDNA and EST. As per AtGDB, the gene structure diagram consists of gene structure representations in which rectangular boxes are used to show exons, lines connecting these boxes depict introns, and arrowheads imply forward and reverse strand transcription. GenBank supplied gene structure predictions are shown as dark blue arrow diagrams. Red arrow diagrams represent the spliced alignment of ESTs. Light blue arrow diagrams are reserved for spliced alignment of known full-length cDNA. Each row also contains a graphic flag surrounded by a red box as used at AtGDB to indicate degree of support for the gene structure annotation. **Row 1** shows the least confirmed level of gene structure annotation in which no known EST or cDNA from this predicted gene exists. 7108 (24.5%) gene annotations within the current *Arabidopsis thaliana* pseudo-chromosome records fall into this category. **Row 2** shows the next level of confirmation in which an EST or cDNA sequence is shown for the region however no splice sites could be confirmed. This level of confirmation implies the existence of a gene, yet tells little of its gene structure. 1119 (3.8%) examples were noted. **Row 3** depicts a case of considerable improvement in confirmation of the given gene structure. These cases include annotations in which at least one splice site is confirmed by EST or cDNA spliced alignment. 3612 (12.4%) annotated genes fit this description. **Rows 4 and 5** are reserved for annotations in which all splice sites are confirmed. Level 4 annotations differ from level 5 only in their sequence coverage. As shown level 4 annotations, 174 (0.6%) cases, include gaps in their sequence coverage whereas level 5 annotations, 16939 (58.5%) cases, are completely covered from first exon to last.

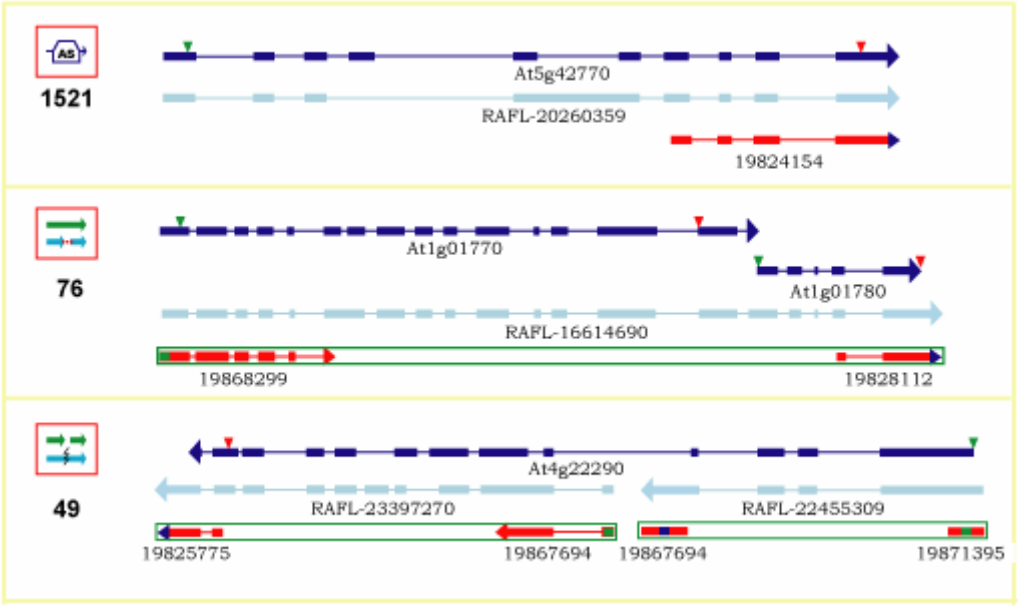

Figure 2. Automated selection of problematic annotations. Annotation flags are used to highlight annotations incongruent with spliced sequences. As in Fig. 1, these flags are depicted with relevant examples and the number of automated finds for each event. Symbols and colors are as described in the legend to Fig. 1. **Row 1** depicts a case of alternative splicing in which the annotated gene structure differs from that evidenced by the full-length cDNA (gi 20260359). 1521 such cases exist for the current *Arabidopsis* annotations. **Row 2** shows evidence of the false prediction of an intergenic region, thus necessitating the union of two adjacent gene structure predictions. 76 other such cases can be found. **Row 3** represents the false prediction of an intron that causes the inaccurate union of two independent gene structures. 49 such cases exist.

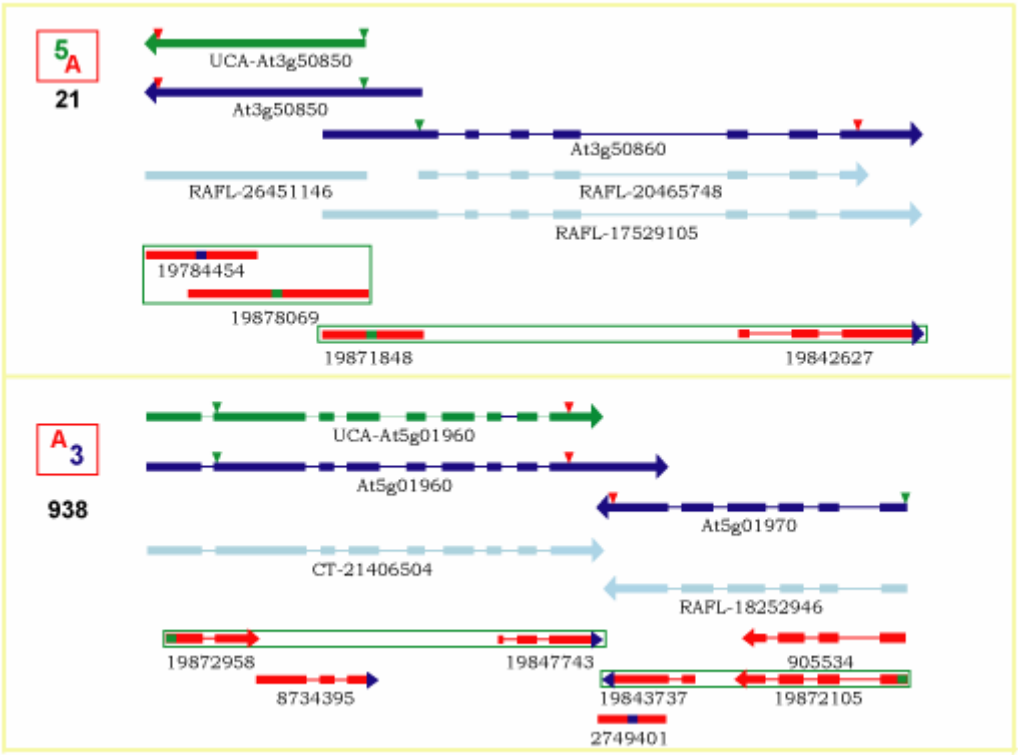

Figure 3. Erroneous assignment of 5' and 3' gene ends. Symbols and colors are as described in the legend to Fig. 1. Green user-contributed gene annotations represent the corrected gene structure supported by assignment of spliced aligned sequences. **Row 1** shows the upstream extension of gene At3g50850 due to the inclusion of EST gi-19871848. This EST clearly originates from gene At3g50860 as evidenced by its green bounding box with neighboring EST gi-19842627, representing the fact that these ESTs are the 5' and 3' respective ends of a single cDNA clone (clone-pair ESTs). **Row 2** shows a similar downstream extension of the gene At5g01960 due to EST gi-2749401.

APPENDIX

The GAEVAL analysis tool

The GAEVAL analysis tool was developed using object-oriented Perl. Perl modules encompass the procedures for comparing isoform specific EST and cDNA spliced (ISE) alignments with individual gene structure annotations. Data from these procedures is stored in a series of relational database tables with MySQL used as the relational database manager. All procedures are generalized and can be adapted to accept alignment and annotation data from numerous sources.

Relational database storage of the data generated during ISE alignment, integrity scoring, and the results of incongruence analysis provide the necessary information for dynamic queries of the annotation set based on supporting evidence. For instance, query reports can be generated which detail the number of annotations with greater than 75% intron confirmation, greater than 50% sequence coverage, which show evidence of alternative splicing, and show evidence of alternative cleavage/polyadenylation site utilization. The sequences of these annotations, or perhaps the genomic sequence proximal to these annotations, can then be retrieved through interaction with a genomic data management tool such as those available at AtGDB (http://www.plantgdb.org/AtGDB/) or OsGDB (http://www.plantgdb.org/OsGDB/).

Scoring gene structure annotations

Gene structure annotation integrity is evaluated by estimating the level of annotation support which is required to provide a structure unlikely to be significantly altered by

reannotation. Annotation support is established through the evaluation of ISEs. Individual elements of the annotation structure (i.e. introns, exons, and UTRs) are compared with each ISE spliced alignment. Verification of the individual structural elements is used to establish an integrity score. The integrity score 'Φ' is given by the formula $\Phi = (.6 * \alpha) + (.3 * \beta) + (.05 * \gamma) + (.05 * \delta)$ where 'α' is the percentage of confirmed intron structures, 'β' represents the percentage of the annotated structure overlapped by at least one ISE, 'γ' the probability density of the 5` UTR length determined by the discrete distribution of observed 5` UTR lengths, and 'δ' the probability density of the 3` UTR length. Weights in the above formula are required to sum to 1 and all parameters are bounded between 0 and 1 thereby normalizing 'Φ'. A substantial number of gene annotations represent intronless transcripts. For these annotations 'Φ' is calculated as above with 'α' determined by the probability density underlying the discrete distribution of observed CDS lengths.

Separate measures are required to ascertain the reliability and completeness of gene annotations in a given transcriptional region. Integrity scoring addresses solely the support of the current annotation. Higher integrity scores denote a more reliable gene structure annotation. Classification of incongruence differs from the scoring of annotation integrity in that incongruence classes may indicate the need for additional transcript isoform annotations and therefore more accurately represents the completeness of the annotation set for a gene region. For this reason, indicators of incongruence are not used as negative values in the derivation of integrity scores

Indication of incongruence

Annotation incongruence is determined by comparison of the annotated gene structure with all ISEs of cognate and non-cognate expressed sequence alignments independently. This comparison determines support for predicted splice junctions, the extent of the annotation region, and the presence of alternative splice junctions. Structure similarities determined during the process of deriving the ISE alignment are used to document the number of ISEs supporting individual exon and intron boundaries in an annotation. Gene annotation incongruence caused by the use of non-cognate (parologous) expressed sequences in determining the annotated gene structure is exposed at this stage. In addition, the extent of the annotated gene, which is often underestimated due to the absence of the codon periodicity in the UTR, is assessed.

Incongruence is made obvious to the GAEVAL algorithm in much the same way as it is visualized at AtGDB (http://www.plantgdb.org/AtGDB/). When the coordinates of intron or exon boundaries derived for an ISE are not equivalent to the overlapping coordinates of intron or exon boundaries given in the gene annotation, incongruence is noted. ISEs derived from both cognate and non-cognate spliced alignments are evaluated. Incongruence noted through comparison with cognate ISEs which appear to be correct when evaluated with non-cognate ISEs are thus determined and provide explanation as to the source of possible error in the annotation.

To determine a source of annotation incongruence and establish an incongruence classification system, individual incongruence indicators are evaluated. A scoring function is applied to each indicator to estimate a confidence value for its association with the given gene annotation. The general form of this function, $\theta = \sum(\alpha_i * A_i)$, assigns

the confidence score ‘θ’ as the sum over all indicators ‘i’ of the product of the indicator influence ‘α_i’ and the ratio of the indicator observance ‘A_i’. Indicator influence provides a parameter to establish the weight of one property indicator relative to another in establishing confidence of a given type of incongruence. While most annotations show significance to one or no incongruence indicator, these indicators are not mutually exclusive. Annotations located in regions of high transcriptional complexity may produce significant confidence scores for more than one type of incongruence.

The impact of an individual case of incongruence will largely depend on the specific use of the incongruent gene annotation. For example, incomplete UTR definitions within a gene annotation, the most prevalent form of incongruence, may have much less impact on the functional annotation of a gene than on the analysis of its transcriptional and translational regulation. Of the more problematic classes of incongruence, complete structural inconsistencies will exist for any annotation collection yet is generally limited to only a few failures. More often, incongruence can be classified into the cases of incomplete UTR definition, alternative splicing and complex transcriptional processing.

CHAPTER 4: TSiP: TRANSCRIPTIONAL START SITE IDENTIFICATION IN PLANTS

A paper to be submitted to *Genome Research*

Shannon D. Schlueter[1]

ABSTRACT

With the sequencing of multiple plant genomes underway, the problem of promoter prediction in plant genomic sequence has taken on significant practical importance. Herein, I introduce a general probabilistic model of plant promoters and the application of this model to promoter prediction in the genomic sequence of *Arabidopsis thaliana* and *Oryza sativa*. To evaluate the capabilities of a model trained solely on plant sequences, I compare the predictive performance of a program called TSiP which applies this model with other notable promoter prediction systems.

INTRODUCTION

The recognition of eukaryotic promoters in genomic DNA sequences by computational analysis is a notoriously difficult problem. Previous studies have reviewed a number of computational approaches to promoter prediction (Fickett and Hatzigeorgiou 1997; Prestridge 2000). Specifically, they have reported on the accuracy of each method in predicting the transcription start site (TSS). The TSS is often considered the downstream

[1] Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA 50011-3260, USA

boundary of the promoter. Significant advancement in the modeling of eukaryotic promoters is required to make such approaches feasible (Ohler and Niemann 2001). Recently, a number of approaches have shown considerable improvement in predictive accuracy due to algorithm development and a significant increase in available sequence data. Approaches showing the greatest breakthrough in performance have utilized a discrepancy in guanine and cytosine nucleotide content in the region near potential TSSs as well as CpG dinucleotide frequency in this region (Davuluri et al. 2001). While significantly improving the accuracy of promoter prediction in human and other vertebrate genomic sequences, these advancements have done little to improve promoter prediction in plants where CpG islands are less prominent (Rombauts et al. 2003). Furthermore, these methods are trained primarily for prediction in vertebrate genomes and perform poorly when applied to plant genomic DNA.

Herein, I introduce a general probabilistic model of plant promoters which incorporates basic transcriptional, translational, and splicing signals as well as compositional features of regions near the TSS. I designed this model to capture the sequence properties of distinct functional units within and surrounding the promoter (e.g. the core promoter, the proximal promoter, and the 5` UTR) without requiring prominent features such as the TATA-box or overrepresentation of common transcription factor binding sites. Detection of initial exons, as is done by First Exon Finder (Davuluri et al. 2001), was integrated into the model. Presence of an intron however is not required to locate a TSS. Model states representative of the TATA-box and TIS are also included yet not required for TSS prediction. The prediction method employed by TSiP is most similar to the scanning Hidden Markov Model approach adopted by the program McPromoter

(Ohler et al. 2000) but differs from this and other programs in several key respects. First, the model structure is non-restrictive. Inclusions of TATA-box, TIS, or donor site predictions are used to provide biologically consistent model structure however are not required to generate a valid TSS prediction. Second, TSiP is trained specifically for plant promoter prediction and makes no model assumptions based on the presence of CpG islands or regions of nucleotide composition strand bias.

RESULTS

To evaluate the current state of computational promoter prediction in plants, I analyzed the relative performance of TSiP and three notable promoter prediction systems: Eponine (Down and Hubbard 2002), First Exon Finder (Davuluri et al. 2001) and TSSP (Shahmuradov et al. 2003) using three plant sequence datasets (see Methods). Each program was executed with default parameter settings as stated in its most recent publication. Relative performance indicators were based on the count of true positive, false positive and false negative predictions. Using the criterion established by Fickett and Hatzigeorgiou (1997), I considered TSS predictions true positives if a prediction falls within the region [-200, +100] relative to a full-length cDNA (fl-cDNA) mapped TSS. TSS predictions occurring outside this range are considered false positives. Mapped TSSs with no predictions occurring within the [-200, +100] range are counted as false negatives.

I first evaluated the predictive performance of each program using an *Arabidopsis* promoter dataset of 10,736 sequences derived from the region [-500, +500] relative to mapped TSS positions of fl-cDNA confirmed gene annotations. The results of this

evaluation are summarized in Table 1. For each sequence, only the highest scoring TSS prediction was considered. This evaluation is expected to give optimal measures of specificity for each program by limiting the occurrence of false-positive predictions. These results are comparable with what may be expected by combining each promoter prediction program with current methods of gene structure prediction.

Both TSiP and TSSP perform extremely well with sensitivity values of 94% and 72% respectively. This is to be expected as both utilize *Arabidopsis* sequence data in their modeling. Interestingly however, the predictive approach of TSSP differs considerably from TSiP suggesting that integrating the results of each may lead to better overall performance. By comparison, the Eponine and First Exon Finder (FirstEF) predictions are largely uninformative. However, Eponine does have the distinction of being the only non-plant associated prediction method to correctly predict more than 100 (1%) of the promoter region TSSs using its default parameter settings. To demonstrate the effect of parameter selection on this evaluation, I significantly lowered the threshold parameters of both Eponine and FirstEF. These results are shown in parenthesis in Table 1. FirstEF was used with the minimal value of 0.2 for all three parameter options while Eponine was limited by a cutoff of 0.4. While bringing the number of true positive predictions closer to that of TSiP and TSSP, these parameter selections drastically increase the number of false positives detected by each program. Note that the counting procedure used to analyze this dataset utilized the highest scoring TSS site of each sequence. As such, each sequence was considered only once as either a true positive, false positive, or false negative. This procedure inadvertently limited the impact of sequences with multiple false positive predictions.

I next evaluated each program using a dataset of 100 *Arabidopsis* genomic sequence segments, each of length 150-200Kb containing between 25 and 42 (average 31) fl-cDNA confirmed gene annotations. This dataset was used to analyze the performance of each program when faced with detecting TSSs in large segments of genomic DNA as would be generated by large scale sequencing endeavors. The results of this evaluation are summarized in Table 2. For this evaluation less stringent Eponine and FirstEF parameter selection was used as before. As expected, the ratio of false-positive to true-positive predictions is substantially higher for all programs tests. Each TSS prediction in this evaluation was independently counted (not accounting for output binning performed by each program) showing the full impact of false positives.

In order to provide an unbiased evaluation, I concluded these tests by evaluating each program using a dataset of 50 *Oryza sativa* (rice) genomic sequence segments. Each test sequence is of length 150-200Kb and contains between 22 and 36 (average 27) fl-cDNA confirmed gene annotations. The results of this evaluation are summarized in Table 3. While the predictive performance of TSiP is lower in this evaluation than that of the *Arabidopsis* dataset, its relative performance as compared with the other tested methods is a clear indicator of the effect of training data on model performance.

DISCUSSION

As the capacity of high-throughput sequencing advances, large genomic sequence collections of numerous plant species will become available. The need for accurate methods to identify biologically significant regions in these anonymous genomic sequences is paramount. Experimental investigation will continue to be essential to verify

the exact location and characterize the functional significance of these regions; however bioinformatic approaches have already shown promise in reducing this enormous search space to a manageable size. Herein, I have described the development of a probabilistic model of promoter structure and its application to TSS prediction in two plant species.

Owing to the absence of prominent CpG islands in plant genomic sequence, many of the methods currently available for promoter prediction are ineffective. Efforts to parameterize these methods for better performance on plant sequence data showed little promise (Rombauts et al. 2003). Sequence signals such as GC compositional strand bias or GC-skew (=(C-G)/(C+G)) have recently been suggested as indicators of TSS position (Fujimori et al. 2005) which may be used in place of CpG islands. As shown in figure 1A, the averaged base composition taken over a sufficiently large collection of sequences does show a significant trend toward increased cytosine levels in the immediate upstream region of the TSS. This appears to be a causative source of the observed GC-skew for these regions (figure 1B). Individual profiles however show a high degree of variation when measuring GC-skew thus limiting its effectiveness as a TSS indicator (figure 2). The presence of the subtle GC-skew trend (which is not seen in vertebrate promoter sequences) does indicate a potential for probabilistic differentiation of this region from surrounding sequences and a necessity for modeling data based solely on plant sequence collections.

The model architecture I chose for capturing the functional units of plant promoters is admittedly more accurate in selecting probable parses of the region downstream of the TSS (the transcript region). This is largely due to the presence of signal sites delineating the boundaries of intervening sequence regions in this part of the

model (see Methods). In the region upstream of the TSS (the promoter proper), boundaries are less clearly defined and in fact distinctions were not previously annotated in the case of proximal promoters versus core promoters. The distinction between these regions necessitated the use of a core promoter model based on a 4^{th} order rationally interpolated markov chain of the region 50nt upstream of the annotated TSS. Areas of the promoter in which this model produced log odds scores significantly less than a null model based on nucleotide frequencies of the promoter region ([-500, -1] relative to the TSS) were selected as examples of proximal promoter states. In figure 3, I display a plot of log odds ratios for the proximal and core promoter states in correlation with transcription factor binding sites predicted by the TSSP program. The correlation between transcription factor binding site motif alignments generated by TSSP and the prevalence of the proximal promoter state model over the core promoter model is currently being investigated.

The TSiP probability model allows for the calculation of conditional probability scores for both optimal and suboptimal TSS predictions. This may be of interest in both the context of transcriptional initiation efficiency and alternative TSS usage. Promoters characterized by a single TSS show a single narrow distribution of approximately 50 nucleotides which peaks above a conditional probability score of 0.15 (figure 4). With this, I have contrasted three previously reported promoters characterized by multiple TSSs. These comparisons show an interesting application of TSiP in the functional characterization of individual promoters.

The At1g76490 *Arabidopsis thaliana* gene locus was previously reported to encode a 3-hydroxy-3-methylglutaryl coenzyme A reductase gene which made use of an

alternative promoter to encode an N-terminal extended isoform (Lumbreras et al., 1995). The conditional probability plot of the TSS state based on the TSiP model for this region shows a very broad distribution of 178 nucleotides with relatively weak scores (figure 5). The primary peak occurring at nucleotide position 28700640 corresponds to the standard transcript isoform. The secondary isoform can be seen occurring near the 28700740 nucleotide position. A bimodal distribution applied to this plot confirms both TSS locations as well as their relative transcriptional efficiencies.

The At5g47770 gene locus was previously reported to encode a farnesyl-diphosphate synthase gene which generates a novel transcript isoform targeted to plant mitochondria (Cunillera et al., 1997). The plot for this region confirms the presence of a downstream alternative TSS (figure 6). Notice the relatively weak conditional probability scores due to the presence of both TSSs within the parse region and the altered gene structure of the secondary isoform. Interestingly the optimal TSiP model parse associated with each TSS location accurately identifies the first intron donor site depicted in the figure.

The At1g63940 gene locus was previously reported to encode a monodehydroascorbate reductase gene which uses multiple TSSs to provide dual targeting of the protein to both chloroplasts and mitochondria (Obara et al., 2002). Like the previous examples, the plot of conditional probability scores for the TSS state base on the optimal TSiP model parse confirm the presence of multiple TSS locations (figure 7). Interestingly, the two TSS sites previously reported are the two downstream sites whereas a third site located at nucleotide position 23733700 was not mentioned and appears to be associated with an alternatively spliced transcript.

I have shown that probabilistic modeling can provide preliminary evidence as to the location and structure of plant promoters. By training such models on verified sequences from a single plant species, I can report improved performance when compared with existing methods. I believe the source of this performance is due primarily to the sole use of plant sequences for training. I anticipate further refinements to the model architecture such as bi-directional transitions between the proximal and core promoter states and introduction of filtering approaches such as a support vector machine to integrate probability measures from non-adjacent states. I enthusiastically endorse the use of multiple analysis methods for plant promoter identification as no single model will be significantly complex to provide 100% accuracy in the near future.

METHODS

Arabidopsis thaliana promoter sequence training set

As a primary set of positive examples for use in model development, I extracted genomic sequence regions and gene annotations from the *Arabidopsis* version 6 chromosome pseudo-molecules (accessions: NC_003070.5, NC_003071.3, NC_003074.4, NC_003075.3, and NC_003076.4). Extracted regions extended from 1000nt upstream of an annotated TSS to 100nt downstream of the closer of either the translation initiation site (TIS) or the first intron donor site. From AtGDB, I obtained the spliced alignments of 688233 *Arabidopsis thaliana* expressed sequence tag (EST) and full-length cDNA (fl-cDNA) sequences. Using the GAEVAL analysis package (Schlueter *unpublished*). I then filtered, trimmed, and annotated the extracted genomic sequences using the following criteria. I removed sequences if there were no fl-cDNA alignment supporting the

annotated TSS within a 50nt window. When an adjacent upstream annotation was within 2000nt, I trimmed the leading sequence such that it would contain only half the intergenic region. If trimming resulted in less than 300nt between the beginning of the sequence and the annotated TSS, I removed the sequence. If GAEVAL analysis detected an incongruent or alternative intron in the specified region, I discarded the sequence. Finally, I annotated each sequence by documenting the specific state (intergenic, TSS, 5`UTR, TIS, coding, intron donor, intron) of each nucleotide.

In total, I obtained 13836 sequences by this procedure. To develop an independent dataset for testing, I removed 3100 sequences which were contained in 100 non-overlapping genomic segments described below. The remaining 10736 sequences exhibited the following properties. 1647 (15%) were derived from intronless gene annotations leaving 9089 (85%) associated with multi-exon gene annotations. Of these, 2329 (26% of multi-exon annotations; 22% of all annotations) are derived from annotations with completely non-coding first exons leaving 6760 (74% multi-exon; 63% all annotations) with portions of their first exon involved in translation. Each sequence was further classified by the presence or absence of a TATA-box signal in the region 15 to 60 nucleotides upstream of the annotated TSS. TATA-box signals were assigned as described below. In total, 1318 (12%) of the sequences in this dataset were determined to be strongly associated with a TATA-box.

Arabidopsis thaliana genomic sequence test set

As an initial set of test sequences, I choose 100 genomic sequence segments of length 150-200Kbps. Using the GAEVAL analysis package, I selected sequences such that they

were of the prerequisite length and were verified to contain solely annotations which passed the filtering criteria described above. Furthermore, I removed from consideration any sequence segment which contained EST or fl-cDNA spliced alignments not associated with a verified annotation. The resulting test sequences contained an average of 31 gene annotations (range: 25-42) per sequence segment. My intention for this test set is to provide a collection of large genomic segments with controlled annotation of intergenic regions and TSS locations. I removed the resulting 3100 sequences associated with this test set from the *Arabidopsis* promoter training set to prevent bias.

Oryza sativa genomic sequence test set

As a secondary set of test sequences, I choose 50 genomic sequence segments from the pseudo-molecule assemblies of *Oryza sativa* (TIGR version 4). Sequences were chosen using the same criteria as described for the *Arabidopsis* datasets. Spliced alignments of EST and fl-cDNA sequences used in this evaluation were obtained from OsGDB (http://www.plantgdb.org/OsGDB/). The resulting test sequences contained on average 27 verified gene annotations (range: 22-36).

Model of promoter and TSS regions

Figure 8 illustrates a general model of the genomic structure surrounding the TSS. Individual states of this model correspond to functional genomic sites (represented as circles) and genomic regions (represented as rectangles) which may occur in any biologically consistent order. Arrows representing transitions from one state to the next depict implied biological constraints. Note that with the exception of the transitions from

the intergenic state to the proximal promoter state and the proximal promoter state to the core promoter state, all states transition between regions and sites (rectangle to circle; circle to rectangle). Though similar to the stochastic segment model described by Ohler *et al.* (2001), this model is non-restrictive in that it includes both TATA and TATA-less promoters as well as structure consistent with both intron-containing and intronless transcripts. This model, similar in design to the general Hidden Markov Models used by GENMARK (Lukashin and Borodovsky 1998), GENSCAN (Burge and Karlin 1997), and McPromoter (Ohler and Niemann 2001), derives from earlier works in the field of speech recognition (Rabiner 1986) and probabilistic gene prediction (Krogh 1997). Probability generating models correspond to each state type were generated as described below. Following the convention of Burge and Karlin (1997), transition probabilities and duration distributions were estimated from the annotated training dataset.

Signal site models

Numerous options exist for modeling biological signal sites. For the site based states of this model, I chose to use a generalization of the weight array model. The weight array model itself is a generalization of the weight matrix or profile method in which the nucleotide frequencies at each position of the profile model are replaced with a nucleotide generating probability conditioned on the previous nucleotide of the matrix (Zhang and Marr 1993). The method I use in modeling the TSS, TATA, TIS, and donor sites is to construct a 2^{nd} order weight array model (W^2AM) of each in which nucleotide generating probabilities are conditioned on the previous dinucleotide. The TIS and donor

site W^2AM models are each 40nt models with the relevant site located at position 21. The TATA and TSS models however are 30nt models with sites located at position 11.

Signal region models

For the region based states of this model, I have used interpolated markov chains (Ohler et al. 1999) with orders corresponding to the average size of the region. I applied a 7^{th} order rational IMC to the intergenic, proximal promoter, and intronic regions, a 5^{th} order rational IMC to the intronless and intron-containing coding regions, and a 4^{th} order rational IMC to the remaining region models.

TATA signal training and assignment

Using MotifScanner (Thijs et al. 2001) and a previously described *Arabidopsis thaliana* TATA-box motif (Molina and Grotewold 2005), I selected sequences from the *Arabidopsis* promoter dataset exhibiting motif matches in the region 15 to 60 nucleotides upstream of the annotated TSS. 1031 sequences were detected by this procedure. Using the motif position as an anchor, I then generated a 30nt W^2AM model from these sequences. I annotated the promoter dataset using this model with the [-60, -15] region criterion and a log odds score criterion of 2.

Application of the promoter model

To determine the optimal parse of sequence, I use a standard Viterbi algorithm (Viterbi 1967). I then calculate a score for any TSS state present in the optimal parse. TSS scores are derived from the conditional probability of the TSS state at the position indicated by

the optimal parse given the sequence. This probability can be efficiently calculated by applying the "forward-backward" algorithm (Rabiner 1989). Prediction is preformed by TSiP in one of two modes (analyze or scan). In analyze mode, TSiP applies the promoter model to a 1000nt window at the beginning of the genomic sequence and every 100nt thereafter. TSS scores are tracked and local maxima are reported. This mode is intended for smaller sequences believed to contain a TSS. In scan mode, TSiP applies the promoter model in 1000nt windows positioned by pre-scanned scores of the TATA, TIS, and donor site models. TSiP first determines log odds scores of each site model and then places the window according to a threshold parameter of each model. This mode is intended for use on large genomic sequence scans.

ACKNOWLEDGMENTS

Shannon Schlueter was supported in part by the National Science Foundation Integrative Graduate Education and Research Traineeship (IGERT) grant DGE-9972653.

REFERENCES

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78-94.

Cunillera N, Boronat A, Ferrer A. 1997. The Arabidopsis thaliana FPS1 gene generates a novel mRNA that encodes a mitochondrial farnesyl-diphosphate synthase isoform. *J Biol Chem* **24**: 15381-15388

Davuluri, R.V., I. Grosse, and M.Q. Zhang. 2001. Computational identification of promoters and first exons in the human genome. *Nat Genet* **29:** 412-417.

Down, T.A. and T.J. Hubbard. 2002. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res* **12:** 458-461.

Fickett, J.W. and A.G. Hatzigeorgiou. 1997. Eukaryotic promoter recognition. *Genome Res* **7:** 861-878.

Fujimori, S., T. Washio, and M. Tomita. 2005. GC-compositional strand bias around transcription start sites in plants and fungi. *BMC Genomics* **6:** 26.

Krogh, A. 1997. Two methods for improving performance of an HMM and their application for gene finding. *Proc Int Conf Intell Syst Mol Biol* **5:** 179-186.

Lukashin, A.V. and M. Borodovsky. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26:** 1107-1115.

Lumbreras V, Campos N, Boronat A. 1995. The use of an alternative promoter in the Arabidopsis thaliana HMG1 gene generates an mRNA that encodes a novel 3-hydroxy-3-methylglutaryl coenzyme A reductase isoform with an extended N-terminal region. *Plant J* **4**: 541-549.

Molina, C. and E. Grotewold. 2005. Genome wide analysis of Arabidopsis core promoters. *BMC Genomics* **6:** 25.

Obara K, Sumi K, Fukuda H. 2002. The use of multiple transcription starts causes the dual targeting of Arabidopsis putative monodehydroascorbate reductase to both mitochondria and chloroplasts. *Plant Cell Physiol.* **7**: 697-705.

Ohler, U., S. Harbeck, H. Niemann, E. Noth, and M.G. Reese. 1999. Interpolated markov chains for eukaryotic promoter recognition. *Bioinformatics* **15:** 362-369.

Ohler, U. and H. Niemann. 2001. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet* **17:** 56-60.

Ohler, U., G. Stemmer, S. Harbeck, and H. Niemann. 2000. Stochastic segment models of eukaryotic promoter regions. *Pac Symp Biocomput***:** 380-391.

Prestridge, D.S. 2000. Computer software for eukaryotic promoter analysis. *Methods Mol Biol* **130:** 265-295.

Rabiner, L.R. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE* **77:** 257-285.

Rabiner, L.R.a.J., B.H. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine* **3:** 4-16.

Rombauts, S., K. Florquin, M. Lescot, K. Marchal, P. Rouze, and Y. van de Peer. 2003. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol* **132:** 1162-1176.

Shahmuradov, I.A., A.J. Gammerman, J.M. Hancock, P.M. Bramley, and V.V. Solovyev. 2003. PlantProm: a database of plant promoter sequences. *Nucleic Acids Res* **31:** 114-117.

Thijs, G., M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17:** 1113-1122.

Viterbi, A.J. 1967. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Informat. Theory* **IT-13:** 260-269.

Zhang, M.Q. and T.G. Marr. 1993. A weight array method for splicing signal analysis. *Comput Appl Biosci* **9:** 499-509.

Table 1. **Performance on *Arabidopsis* genomic sequences localized to the promoter region**

Method	# of TSSs	TP	FP	FN	S_n	S_p
TSiP	10736	8697	1521	518	0.9438	0.8511
TSSP	10736	6517	1717	2502	0.7226	0.7915
Eponine	10736	2897 (5689)	2462 (4830)	5377 (217)	0.3501 (0.9633)	0.5406 (0.5408)
FirstEF	10736	82 (3250)	46 (2975)	10608 (4510)	0.0077 (0.4188)	0.6406 (0.5221)

The maximum allowed distance between the predicted TSS and the annotated TSS is 200 nucleotides upstream and 100 nucleotides downstream [-200, + 100].

TP : True Positive, **FP** : False Positive, **FN** : False Negative
$\mathbf{S_n}$ (sensitivity) = TP/(TP+FN), $\mathbf{S_p}$ (specificity) = TP/(TP+FP)

Table 2. **Performance on *Arabidopsis* genomic regions with full-length cDNA confirmed annotation**

Method	# of TSSs	TP	FP	FN	S_n	S_p
TSiP	3100	2372	7588	728	0.7652	0.2382
TSSP	3100	1972	5381	1128	0.6361	0.2682
Eponine	3100	1677	14778	1423	0.5410	0.1019
FirstEF	3100	1821	9251	1279	0.5874	0.1645

The maximum allowed distance between the predicted TSS and the annotated TSS is 200 nucleotides upstream and 100 nucleotides downstream [-200, + 100].

Table 3. **Performance on *Oryza sative* (rice) genomic regions with full-length cDNA confirmed annotation**

Method	# of TSSs	TP	FP	FN	S_n	S_p
TSiP	1350	983	3359	367	0.7281	0.2264
TSSP	1350	873	2657	477	0.6467	0.2473
Eponine	1350	768	7157	582	0.5689	0.0969
FirstEF	1350	724	4033	626	0.5363	0.1522

The maximum allowed distance between the predicted TSS and the annotated TSS is 200 nucleotides upstream and 100 nucleotides downstream [-200, + 100].
* Values in parenthesis are based on a [-500, +200] criterion for TSS localization.

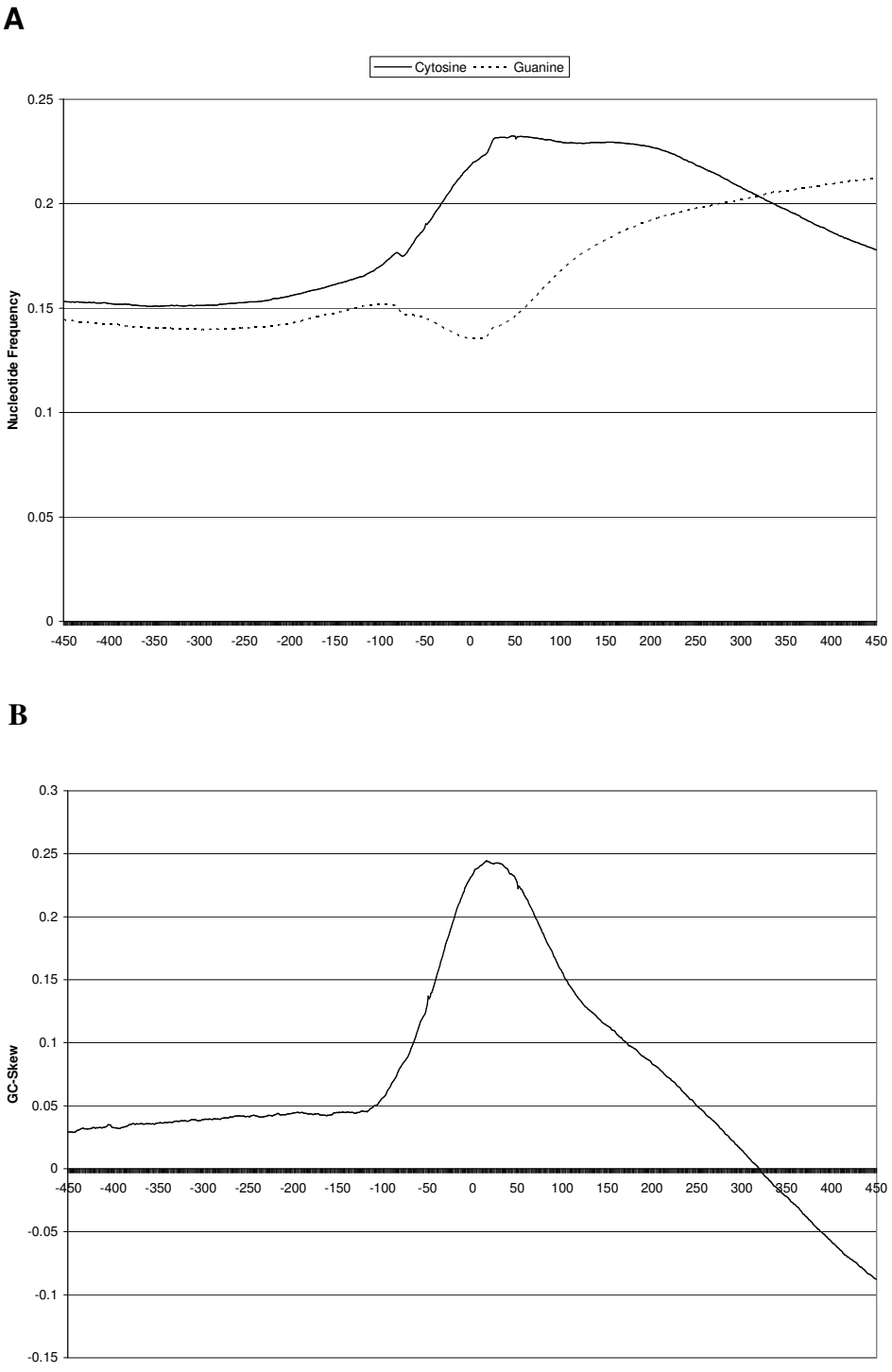

Figure 1. [A] Cytosine and guanine nucleotide frequency in the region [-450, +450] relative the transcription start site of 10,736 *Arabidopsis thaliana* full-length cDNA verified gene annotations. **[B]** Nucleotide composition strand bias or GC-Skew = (C-G)/(C+G).

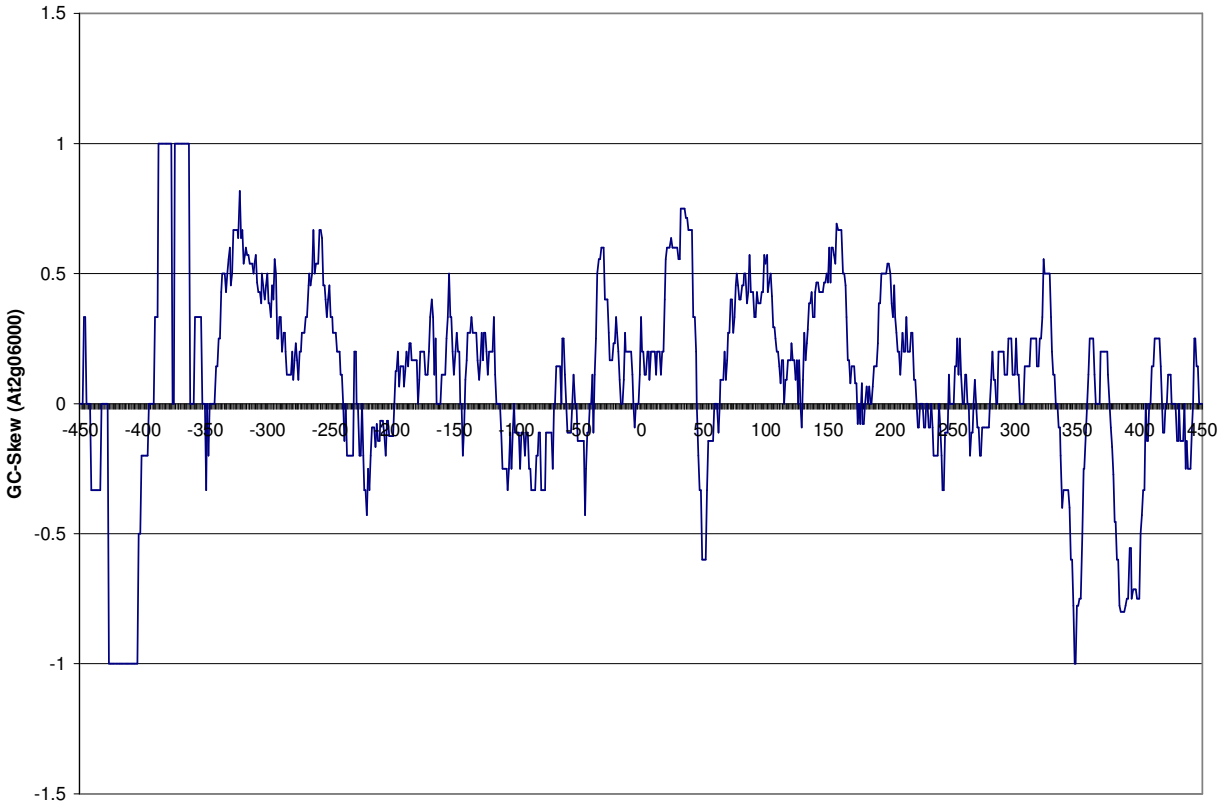

Figure 2. Profile nucleotide composition strand bias for the regions [-450, +450] relative the transcription start site of a single *Arabidopsis thaliana* gene annotation (At2g06000).

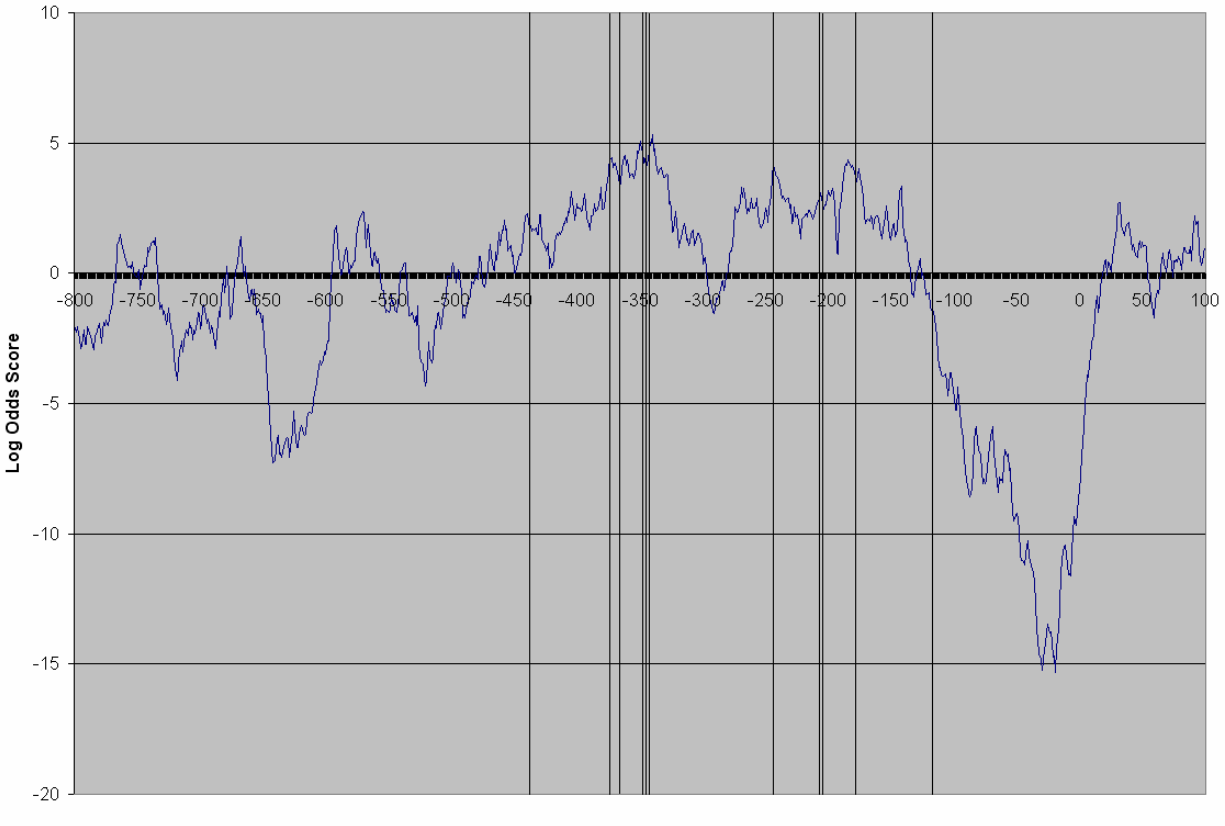

Figure 3. Log odds ratio of Proximal Promoter IMC to Core Promoter IMC.

$$S(x) = \log (P(x|\pi_p) / P(x|\pi_c))$$

Positive values represent preference for Proximal Promoter IMC. Negative values represent preference for Core Promoter IMC. Vertical lines indicate the position of transcript factor binding site motif alignments (excluding TATA related motifs) as predicted by TSSP.

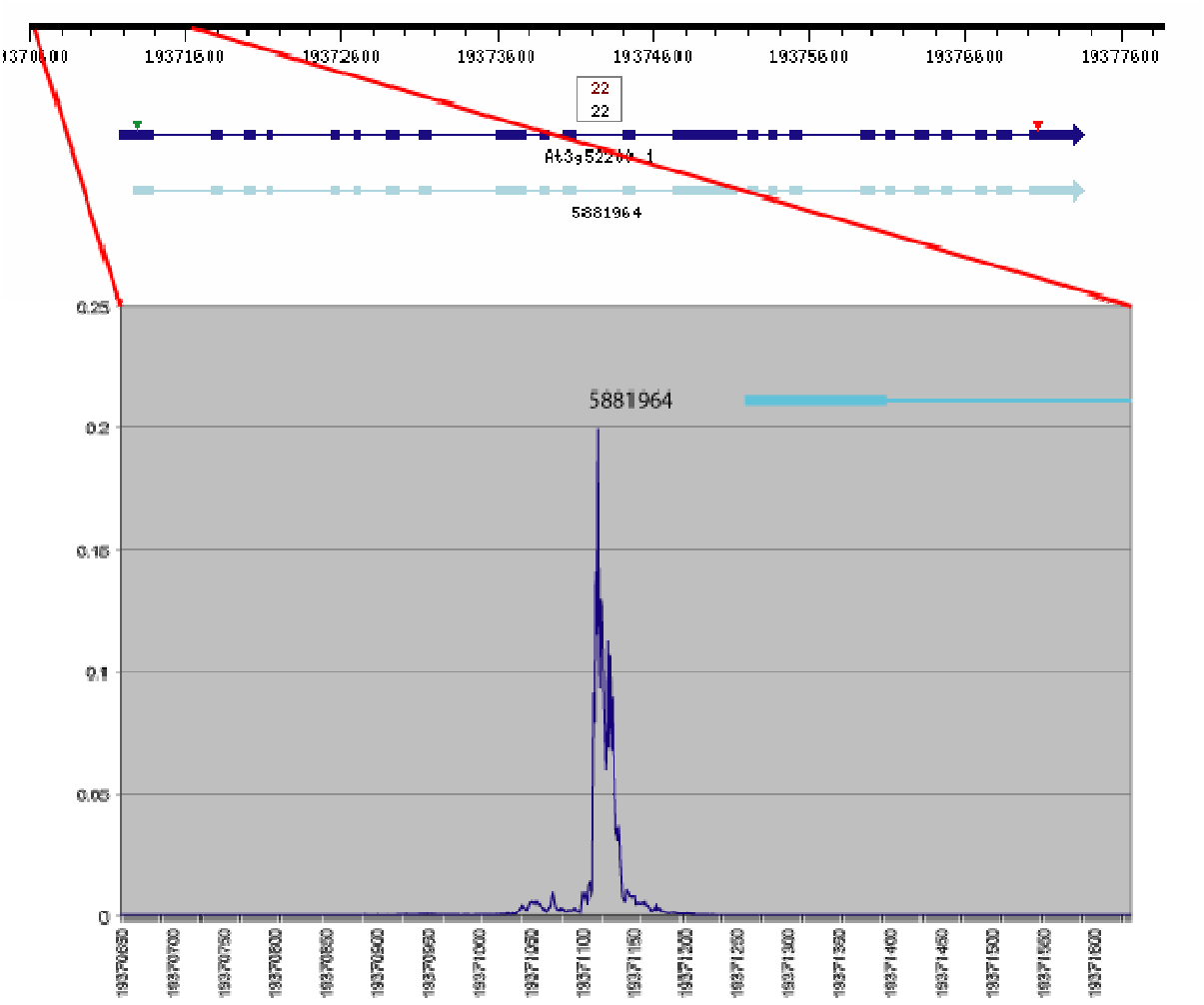

Figure 4. Conditional probability plot of the TSS state for a promoter characterized by a single TSS location. Annotated gene structure for the At3g52200 gene locus is shown in dark blue with rectangular boxes representing annotated exons and lines connecting the boxes representing introns. The green and red triangles above this structure represent the location of translation start and stop codons respectively. The light blue structure represents the spliced alignment of a full-length cDNA.

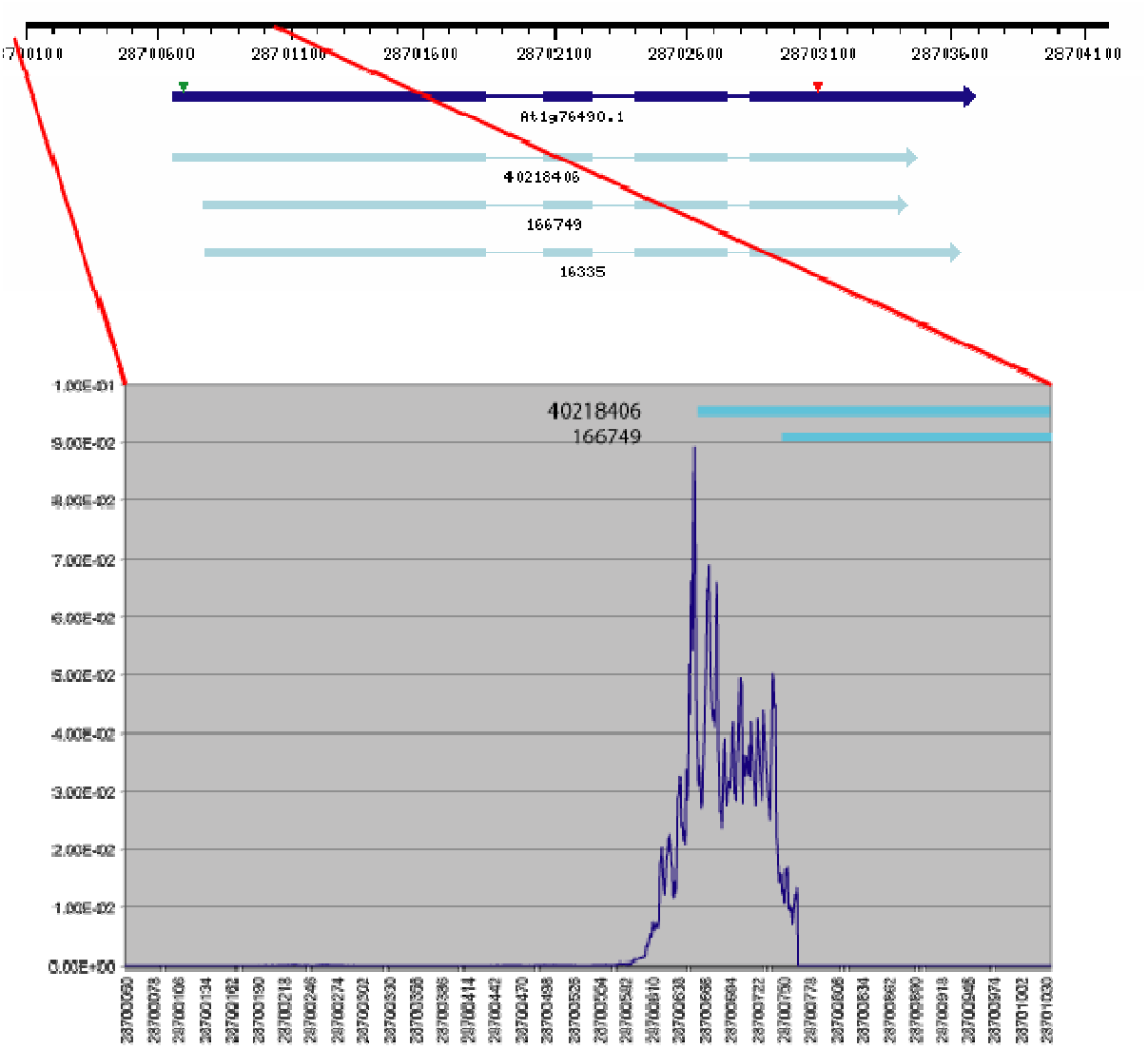

Figure 5. Conditional probability plot of the TSS state for the HMG1 gene locus. Gene annotation structures and cDNA spliced alignment depictions are as described in figure 4.

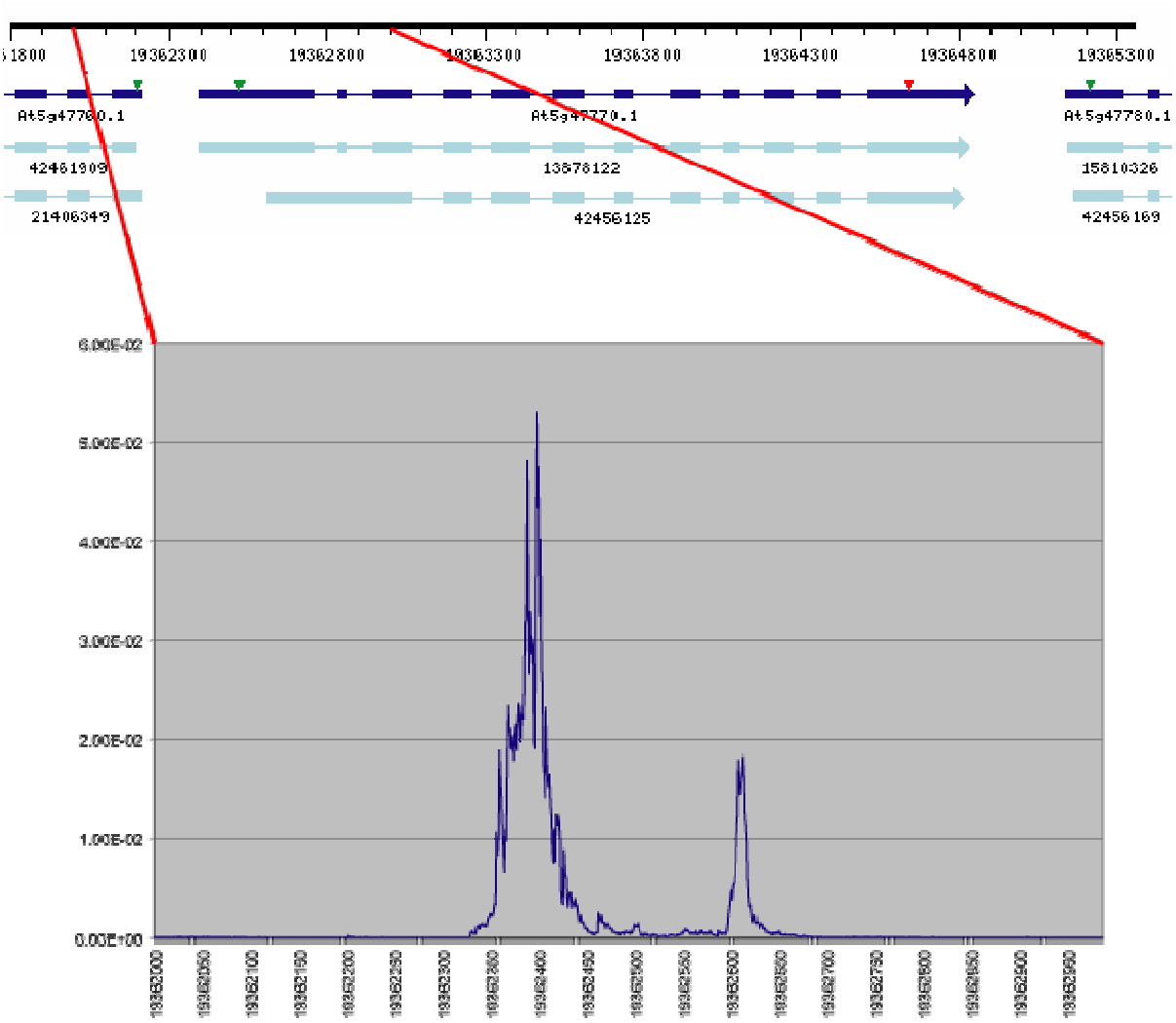

Figure 6. Conditional probability plot of the TSS state for the FPS1 gene locus. Gene annotation structures and cDNA spliced alignment depictions are as described in figure 4.

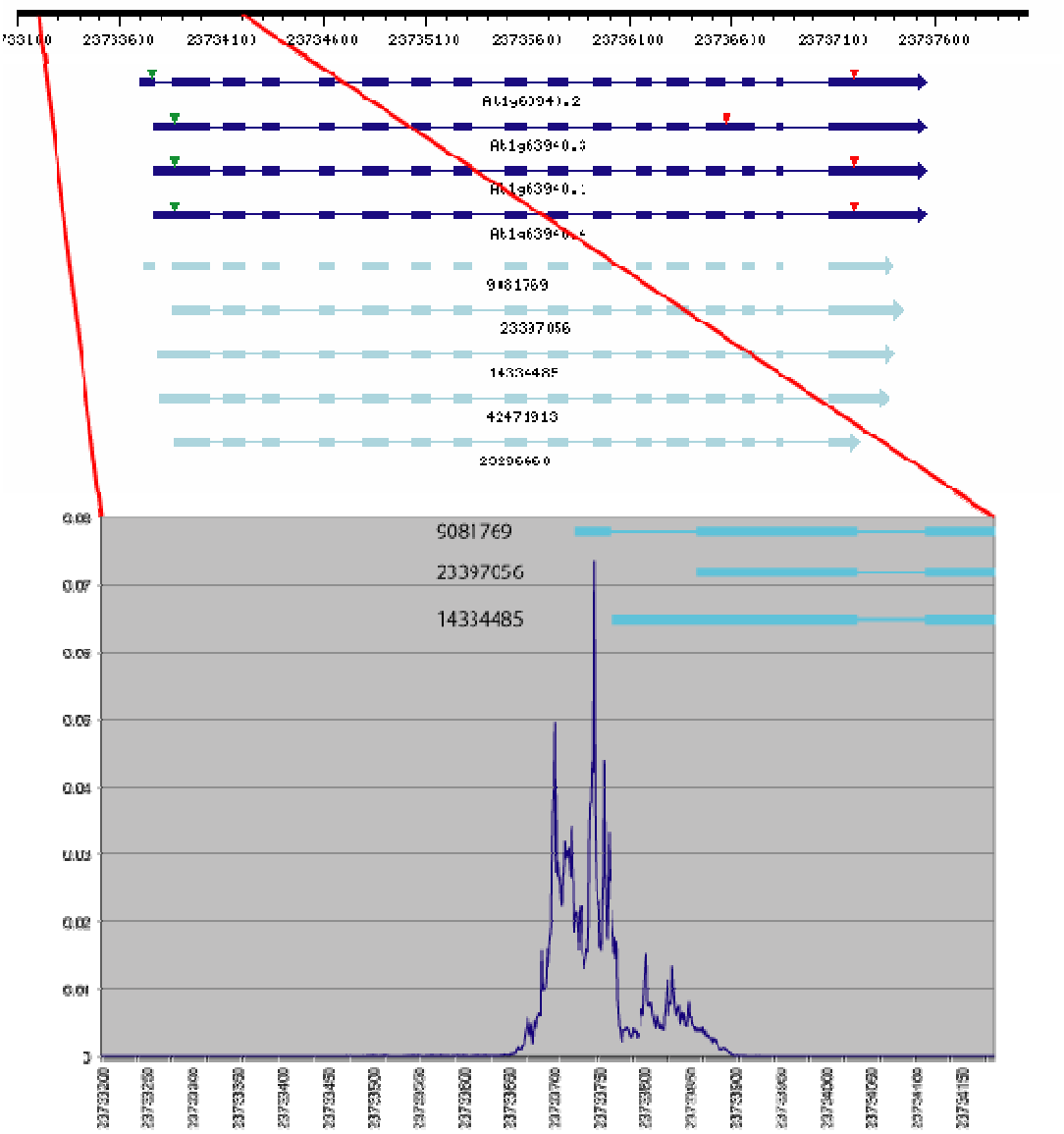

Figure 7. Conditional probability plot of the TSS state for the MDAR gene locus. Gene annotation structures and cDNA spliced alignment depictions are as described in figure 4.

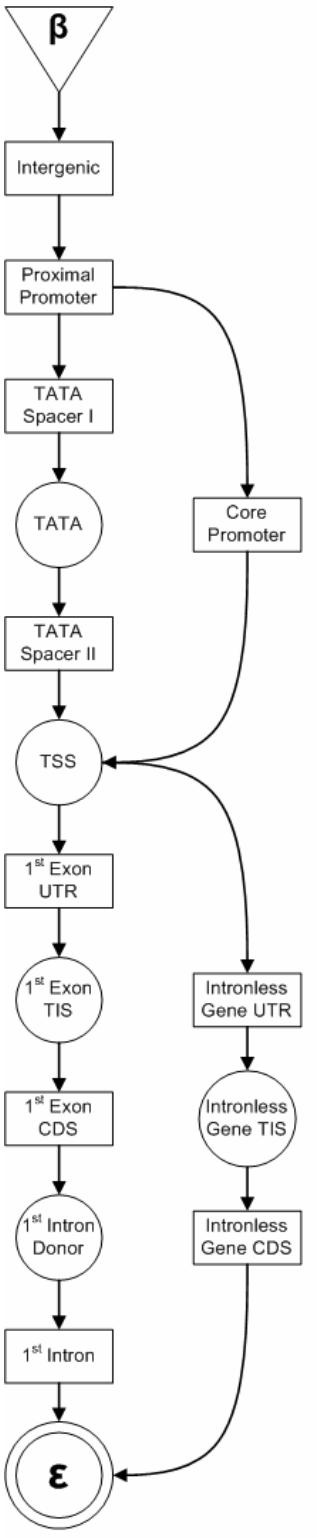

Figure 8. General state model of promoter region.

CHAPTER 5: GENERAL CONCLUSIONS

Conclusions

As biological data collection methods have become more cost effective and less time consuming, the necessity for computational tools to store, manage, and analyze this data has led to the creation of a broad field of research. Bioinformatics, while firmly rooted in the technology of information management, is now a mainstream component in the majority of scientific investigations. With the vast majority of effort in bioinformatics being applied to research on vertebrate species, researchers in the plant sciences have often been left with less than satisfactory tools. The research presented in this dissertation was done in an effort to advance the quality of bioinformatic tools available for plant genomics and to develop a better understanding of the unique aspects of plant biological processes such as the transcriptional processing of protein coding genes.

In the course of this study, I have developed an extensible infrastructure for integrating biological data sources and applying them to hypothesis driven research. Eleven plant species xGDB databases have been made publicly available to facilitate progress in plant genome informatics. A sophisticated system was devised and developed to investigate the reliability of gene structure annotations on a per gene basis. With this, I generated the necessary dataset to develop a plant specific probabilistic model of RNA polymerase II transcription start sites.

The prediction of transcription start sites and promoter regions in plant genomic DNA was found to be considerably more challenging than similar endeavors in vertebrate sequences. Probabilistic models based solely on plant promoter sequences improved the

outlook for promoter prediction in plant genomes. However, owing to the lack of a pervasive signal such as the presence of CpG islands, results are still less than ideal.

In conclusion, progress was made in providing resources tailored to the plant research community and in the investigation of transcriptional processing in plants. Distinct regions which may be functionally significant in the regulation of transcription were discovered. In addition, a number of genes utilizing alternative transcription start sites and alternative cleavage/polyadenylation sites were revealed. The results of this study demonstrate that the process of transcription in plants is significantly distinct from that of other organisms and warrants independent and thorough investigation.

ACKNOWLEDGMENTS

I would like to thank Dr. Volker Brendel for his encouragement, support, and guidance. Without his support I would not be where I am today. I would also like to thank each of my committee members, Dr. Randy C. Shoemaker, Dr. Thomas A. Peterson, Dr. Xun Gu, and Dr. Xiaoqui Huang for their guidance and thoughtful discussion. I am also grateful for the various opportunities presented to me in graduate school through the funding of the NSF-IGERT training fellowship and the ISU Plant Sciences Institute fellowship. Finally, I want to thank my wife Jessica who has been by my side through the sanity and the chaos and my son Harmon for reminding me of what is truly important at the end of the day.