PLAIN Project

The PLAIN Project

The Plant Genomics Data Computational Interface (PLAIN) project is a project to build an application programming interface (API) that will provide high-performance, web-based computational access for plant genomics data. The project will include a new, open technology for generating computational data web services; an open, portable, optimized data warehouse that supports very fast queries of plant biology data; and a plant biology query language, query builder, and query optimizer that will provide a simple way to limit query results to only the required data. The project will provide a range of data access methods to serve the needs of computational biologists and bench biologists for large and custom data sets. Implementation of this software at TAIR will provide REST and SOAP web services for computational use of TAIR data, RSS feeds for TAIR objects, and an implementation of a new query builder for TAIR.

While strong advances have been made in data generation methods including new genome sequencing methods, high throughput phenotyping, protein localization and others, computational access to the resulting data still requires large amounts of both human and machine resources. By addressing this issue through architecture, this project leverages advances in software engineering by combining and applying them to the specific domain of plant genomics. In particular, developing a minimal but effective plant genomics schema using modern data modeling; leveraging model-driven architecture to enable generation of high-performance web services from platform-independent models; and developing a basic, well-formed query language for the plant genomics domain are intellectually challenging tasks that will have a significant impact on the technology required for computational access. Wide adoption of a standard set of web interfaces for computational access to plant genomics resources will greatly simplify the effort required to access and integrate plant genomic data, thereby facilitating computational analyses of the data. By providing an easy, robust, and consistent route to computational data access, standard web interfaces will also facilitate development of new resources that could transform existing datasets and present them in new ways, analogous to mashups using Google Maps data along with real estate listings, weather data, Wikipedia entries, etc. By providing computational APIs and the technological infrastructure to create them as open source tools, this project makes available a key set of technologies to computational biologists beyond TAIR. As the technology proves itself, it can move beyond plant biology into the more general biological realm.

PLAIN Data Warehouse

The PLAIN data warehouse was a research project designed to create a new data model that provides fast and easy access to the data of interest in the TAIR database.

The PLAIN data warehouse is a series of data marts, each focusing on a single central concept and the related information relevant to that concept. Each mart gives the client user the ability to query information about the main concept in a way that optimizes both the amount of effort spent in constructing a query and the processing time required to retrieve the results.

- Shared Resources: This subsystem contains the basic resources shared across all data marts, such as taxon and species variant.
- Locus Detail: This subsystem centers on the concept of locus, a specific location on a chromosome. This data mart contains information on gene models, gene structure, functional annotations, polymorphisms, germplasms, and meta data relevant to all available loci.
- Genomic Region:This subsystem centers on the concepts of a reference genome, a collection of sequences (often chromosomes) taken as the standard for a given organism and genome assembly, and a region in such a genome, a sequence feature with an extent greater than zero. A nucleotide region is composed of bases and a polypeptide region is composed of amino acids. This data mart contains information on genes, transcripts, chromosomes, and contigs, with additional information about polypeptides, CDNAs, and ESTs.
- Protein: This subsystem centers on the concept of a protein, an amino acid possibly related to a transcript. It contains information about the related transcripts, protein domains, and also resource links to additional information about the protein.

Plant/SQL

You can find the code for the project in github. https://github.com/tair/psql

The PLAIN project was funded by the National Science Foundation by the NSF Plant Genomics Data Computational Interface grant.