Machado: a genomic data integration framework for Chado developed with Django

Mauricio de Alvarenga Mudadu, Adhemar Zerlotini Neto

Embrapa Informática Agropecuária

Abstract

Technological advances in biological research has led to a data deluge which have great impact in agriculture, especially in plant and animal breeding. In this regard, genome projects and multiomics experiments generate huge volumes of data that must be stored, mined and transformed into useful knowledge. Furthermore, all this information is supposed to be accessible and, if possible, browseable afterwards. Computational biologists have been dealing with this scenario for over a decade and have been implementing software libraries, toolkits, platforms, and databases to succeed in this matter. Although public wide databases exist, research groups still struggle to store and analyze data with local resources and expertise. The GMOD, or Generic Model Organism Database project, is currently the initiative that made the most advance in producing a "collection of open source software tools for managing, visualizing, storing and disseminating genetic and genomic data". Its biological relational database schema known as Chado is widely adopted and many softwares are able to connect to it. Such softwares usually contain a set of scripts to preprocess the data files in order to provide visualization and search capabilities, but in fact they can only operate independently. The Embrapa's Bioinformatics Multi-user Laboratory have been developing an open source software known as machado, that has a Django model to connect to Chado, thus avoiding extra efforts to make data compatible to the database schema. The machado software has several data loading tools for genomic and transcriptomic data and also for annotation results for tools such as BLAST, InterproScan, OrthoMCL and lsTrap. Machado has also an API to connect to Jbrowse which can be easily setup. A web browsing visualisation tool is implemented using the Apache2 server alongside with Django WSGI library. The haystack software integrated with the elasticsearch engine was used to create an index for querying the data using keywords, identification and annotation entries. Caching is also enabled for fast data retrieving. This project aims to contribute to the research community by producing a modern object-relational framework to store, integrate, query, and visualize all the major genomics data types. Such endeavour takes advantage of the latest Python modules to produce an effective open source resource and facilitate the identification of specific biological components related to economic traits in agriculture. Machado is available at GitHub: https://github.com/lmb-embrapa/machado .

Funding: Embrapa