Bio4j: bigger, faster, leaner

Pablo Pareja-Tobes, Alexey Alekhin, Evdokim Kovach, Marina Manrique, Eduardo Pareja, Raquel Tobes and Eduardo Pareja-Tobes[*]

Oh no sequences! Research Group. Era7 bioinformatics

*eparejatobes@ohnosequences.com

Abstract. Bio4j (http://bio4j.com) is a high-performance cloud-enabled graph-based bioinformatics data platform. It is one of the first and most important graph databases for biological data, specially designed to cope with and manage the huge amount of data brought by NGS technologies: it integrates most data available in UniProt KB (SwissProt + Trembl), Gene Ontology (GO), UniRef (50, 90, 100), RefSeq, NCBI taxonomy, and Expasy Enzyme DBs. Data is organized in a way semantically equivalent to what it represents by taking advantage of the graph structure; in this paradigm it is easy to have many different types of relationships and nodes thus making it perfect for highly interconnected complex data (as it is the case of biological data). From a performance point of view, relational databases with their tabular data structure are not able to respond to some complex queries that are possible to resolve using the graph paradigm; graph databases give you fast local access to all the elements related with each entity, through the edges that connect them with others. In this way, queries which would even be impossible to perform with a standard Relational DB, just take a couple of seconds with Bio4j.

This year has seen important updates and new developments on Bio4j; it now includes 1,216,993,547 relationships and 190,625,351 nodes, close to triplicating the figures from one year ago. We have introduced a new level of abstraction for the domain model, by decoupling the inner database implementation from the relationships among entities themselves. Interfaces has been developed for each node and relationship present in the database, including methods to access both the properties of the entity it represents and utility methods that allow to easily navigate to the entities that will be linked to it.

Implementing that set of interfaces we have developed another layer for the domain model using Blueprints, the de-facto standard for graph data modeling, thus making the domain model independent from the choice of database technology. Building on that, we now offer specifically tuned data binary distributions for TitanDB, yielding a dramatic increase in performance due to vertex-local edge-typed indexes.

The introduction of a module system based in Statika makes now possible to deploy only selected components of the integrated data sets, with Amazon Web Services deployments on hardware specifically configured for them.

Bio4j is open source, available under the AGPLv3 license.

This project is funded in part by the ITN FP7 project INTERCROSSING (Grant 289974).