Bio4J: An Open source biological data integration platform

Pablo Pareja-Tobes[1], Eduardo Pareja-Tobes[1], Marina Manrique[1], Eduardo Pareja[1], Raquel Tobes[1]

[1]Oh no Sequences! , Era7 Bioinformatics , Granada, Spain.

rtobes@era7.com

Abstract. Today's biology involves many times the use of different omics approaches, in particular, data and information from genomes and proteins are frequently difficult to integrate. In order to manage in an integrated way the information about complete genomes and proteomes in BG7 based projects we developed the Bio4J platform (www.bio4j.com). Bio4j is a bioinformatics graph based DB including most data available in UniProt KB (SwissProt + Trembl), Gene Ontology (GO), UniRef (50,90,100), RefSeq, NCBI taxonomy, and Expasy Enzyme DB. The current version of Bio4j (0.7) includes 530.642.683 relationships and 76.071.411 nodes. Bio4j uses Neo4j technology, another Open Source project. Since Bio4j is based on Neo4j graph-based DB it is highly scalable. New data sources and features can be added and what it's more important, the Java API allows you to easily incorporate your own data to Bio4j so you can make the best out of it. Performance is one of the main advantages of the platform. In Bio4j data is organized in a way semantically equivalent to what it represents thanks to the graph structure. That means that queries which would even be impossible to perform with a standard Relational DB, just take a couple of seconds with Bio4j. Bio4j is an open source platform released under AGPLv3. Future developments: data for metacyc are being included in Bio4j. Integration strategies of data from different technologies can take advantage of Bio4J platform since this platform has really integrated data from Uniprot, Genomes (Refseq) and NCBI Taxonomy. Bio4j is freely available.