Bio4j: the bioinformatics data platform

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

Oh no sequences! Research Group. Era7 bioinformatics, Granada, Spain.

What is Bio4j?

Bio4j is a bioinformatics graph-based data platform **integrating** the most representative **open data sources** around **protein information**

Data sources

- **UniProt KB (SwissProt + Trembl)**
- **Gene Ontology (GO)**
- **UniRef (50,90,100)**
- **RefSeq**
- **NCBI Taxonomy**
- **Expasy Enzyme DB**

Data as a service

- Data distribution
- Scalability
- Cost-effectiveness

amazon web services™

Graph databases

- Data is stored in a way that *semantically represents its own structure*
- Incorporating new data is easy ⇒ it's *scalable*
- *Vertex-centric* (local) indexes allow to overcome the supernode problem

It's free and open!

AGPLv3 Free Software

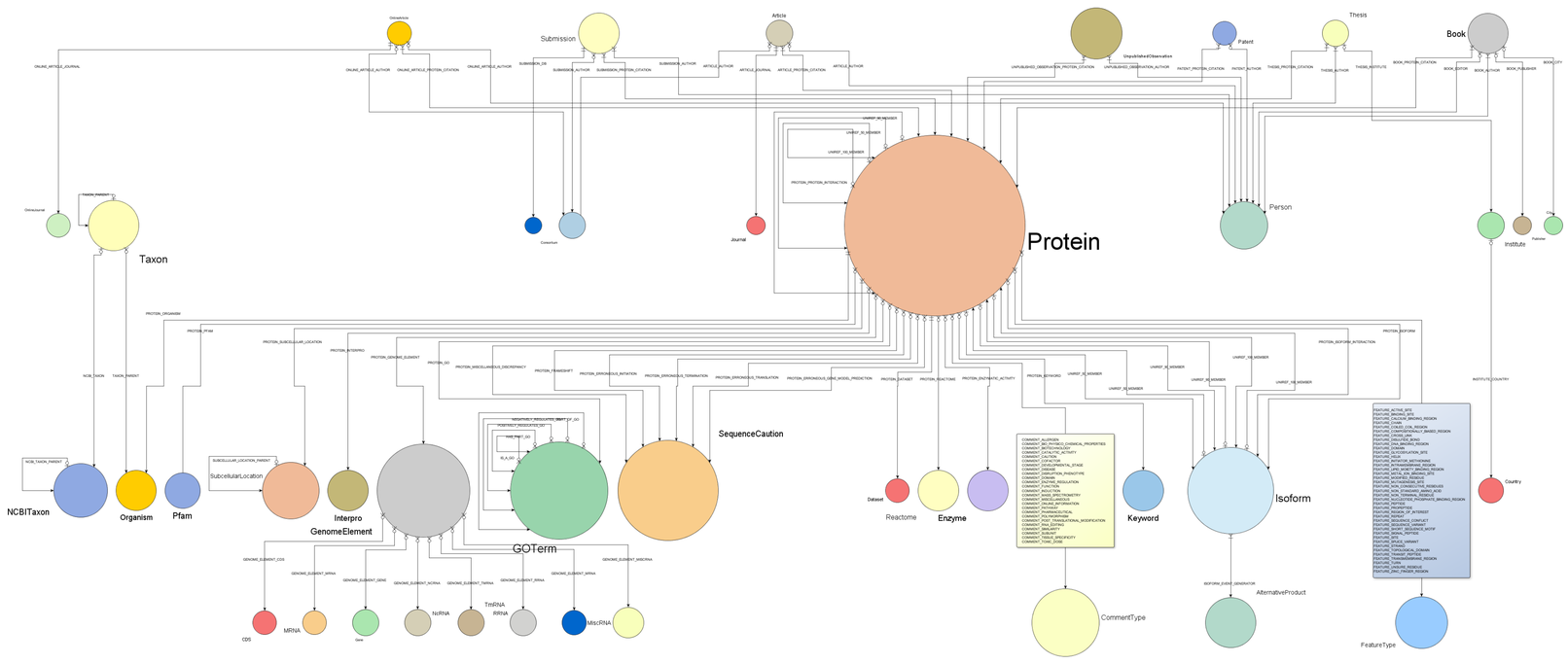
Free as in Freedom

Bio4j is free and open-source under the AGPLv3 license

Bio4j

Developement process is **100% public**. Only **Open Data** is integrated. See

bio4j.com



Bio4j domain model

Bio4j database has a *well-defined* domain model and all nodes and relationships comply with this abstract model.

- 10^9 edges of $\sim$ 150 types
- 2×10^8 nodes of $\sim$ 40 types
- 6×10^8 properties

Different layers of Bio4j

Different *graph topologies* at the storage level, same *domain model* in the client's code.

- Abstract *domain model* with precise typing.
- Universal *Blueprints* implementation.
- *Technology-specific* versions:
  - **Neo4j DB**
  - **TitanDB**
  - **DynamoDB** (WIP)

Use cases

- Era7 Bioinformatics:
  - **BG7** genome annotation
  - **Metapasta** metagenomics analysis
  - Comparative genomics, network analysis, genome assembly
- Ohio State University:
  - Integration and analysis of Chip-seq data
  - Modeling genomic information and gene regulatory networks
- Berkeley Phylogenomics Group:
  - Graph database for Big Data challenges in genomics developed on top of Bio4j

Flexible module system

Statika helps to manage dependencies between modules and simplifies import and deployment in the cloud.
The importing process is *modular* and *customizable* allowing you to import just the data you are interested in.

Some technical details

- Java + Scala source code
- **Statika**-based module system
- **SBT** for building sources and automated tests & release
- **Git + Github**: versioning, docs, collaboration, coordination

Acknowledgments

Bio4j is developed by the R&D team of the Era7 Bioinformatics company

Era7 bioinformatics

ohnosequences!

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

INTER CROSSING

MARIE CURIE ACTIONS