

A Galaxy Interactive Environment for exploring the Neo4j Graph Database

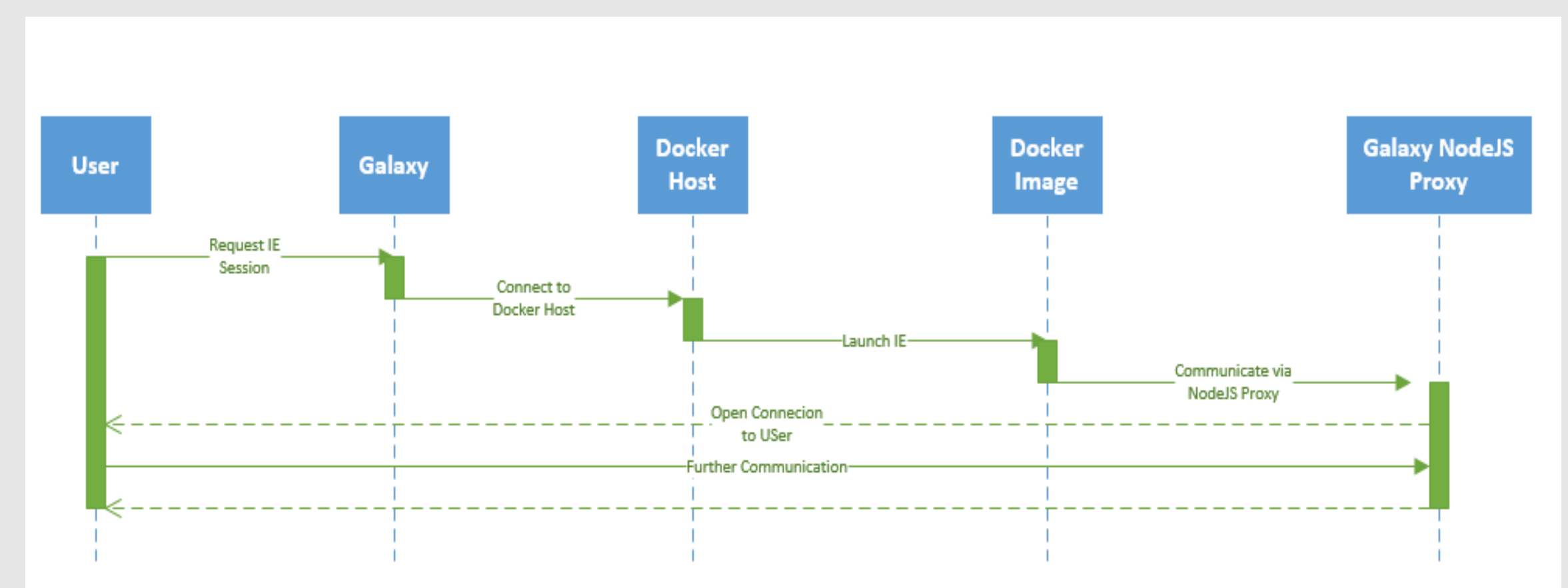
Thoba Lose (thoba@sanbi.ac.za), Peter van Heusden, Alan Christoffels (alan@sanbi.ac.za)
South African National Bioinformatics Institute, University of the Western Cape

Introduction

- Storing biological data involves modeling and storing thousands of entities that are interrelated in complex ways.
- Relational databases meet very specific needs and are not designed to fit all scenarios.
- Graph databases, which focus on connections between entities, are seen as a natural fit to these complex relationships.
- Neo4j is a highly scalable graph database with a declarative query language called Cypher.

Methods

- We implemented a Galaxy Interactive Environment (GIE) to explore a Neo4J database that is stored as a Galaxy dataset.
- The GIE was built by linking a Galaxy plugin to a Docker container based on a modified version of the Neo4J:2.3 Docker image.
- The Interactive Environment is launched from the visualisation menu and is only available for Neo4J database (neostore) datatypes.



Results

The Neo4J database allows for fast querying of thousands of nodes and their relationships, and exploration of the results in a graphical shell. The examples illustrated show the exploration of relationships between two interacting proteins and finding the set of proteins associated with a set of GO terms.



Conclusion

- We have successfully provided an interface to Neo4J databases and the Neo4J graphical shell within Galaxy.
- The nature of the Neo4J shell poses technical challenges for IE implementors. In particular, the shell allows for manipulation as well as querying of the graph database. To overcome these challenges we intend adapting the Neo4J shell to make it more Galaxy-compatible in the future.

The COMBAT TB project is funded by the South African Medical Research Council.

