Integrating disease maps using a graph database approach

Irina Balaur[1], Alexander Mazein[1], Charles Auffray[1]

European Institute for Systems Biology and Medicine (EISBM), Lyon , France[1]

Background: Disease maps are being developed as comprehensive, highly curated and human-readable resources for describing disease mechanisms. The Disease Maps community is continuously extending and currently includes 14 projects (http://disease-maps.org/projects). There is a need for integration of disease maps in a common platform in order to facilitate extension, interrogation and visualization of the integrated data. Graph databases are a natural way to represent and manage biological networks (Lysenko et al., 2016, PMID 27462371; Balaur et al., 2016, PMID 27627442; Toure et al., 2016, PMID 27919219; Balaur et al., 2017, PMID: 27993779; Fabregat et al., 2018, PMID 29377902).

Objectives: We aim to highlight advantages and comment limitations of using the popular graph database Neo4j (neo4j.com) as a core for the common platform for the management (integration, exploration, visualisation) of biological data available in disease maps. Neo4j facilitates network based data integration and provides functionalities for visual exploration of sub networks of interest via a powerful query language (Cypher).

Approach: We discuss several examples where we successfully applied Neo4j for biological knowledge management. Specifically, we first present the Recon2Neo4j framework, which facilitates Recon2 metabolic data exploration using Neo4j (Balaur et al., 2017, PMID: 27993779). Then, we describe a general-purpose UniProt ID-centric framework that has been developed to facilitate exploration of disease context by integrating biological data from major specialized resources on protein-protein interactions, disease-gene associations, drug target relationships, protein-pathway involvement and sequence similarity (Lysenko et al., 2016, PMID 27462371). This resource determined development of specialised Neo4j-based frameworks for asthma and for cardiovascular diseases, to date, by integration of a set of disease-specific implicated genes (denoted here as the “seed genes set”). Finally, we present the STON framework (Toure et al., 2016, PMID 27919219), developed to represent and query information from Systems Biology Graphical Notation (SBGN) diagrams using Neo4j networks.

Conclusion: We anticipate that the use of Neo4j would facilitate quick exploration of the integrated data and identification of common/ overlapping modules within disease maps. A common Neo4j-based framework would offer also the possibility to query all disease maps at once and identify those that include, for example, proteins of interest. However, given the complexity of the integrated maps, the development of the graph data model is not trivial and the effort of the framework implementation has to be well-estimated.

Acknowledgements: The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking the eTRIKS Project (IMI 1154446) resources of which are composed of financial contribution from the European Union’s Seventh Framework Programme (FP7/2007-2013) and EFPIA companies’ in kind contribution.