Abstract: InterMine as a Science Gateway for Systems Biology

Aaron Golden
School of Mathematics, Statistics and Applied Mathematics
National University of Ireland, Galway
Republic of Ireland
aaron.golden@nuigalway.ie

ABSTRACT

Sequencers, protein chromatographers, microarrays, mass spectrometers - all of these devices are now in widespread use in the biomedical sciences, allowing us to perform a complete census of the small molecules within a biological sample - ranging from the genome that ultimately `programs' that sample, to its transcripts, functioning proteins and overall metabolite content. Being able to `link' the resulting diverse datasets in some way offers the possibility of exploring the network of `links' to guide this process of integration, and so consequently, enable a systems biology exploration of the experimental subject [1]. Linking such complex and diverse datasets is best resolved by implementing an ontology, or ontologies, where the properties of a given dataset and metadata are used to create a semantic set of relationships. This resulting semantic ontology can then form the basis for studying relationships and associations between the original data sets, providing a powerful means to analyse the data space in an integrative fashion [2]. The most widely adopted software system that has been developed to date that supports such data integration and exploratory mechanisms is the open-source data warehouse platform InterMine [3]. InterMine uses a core `data model' based around the Sequence Ontology (SO) [4] that provides a powerful means to anchor all possible genomic, epigenomic, proteomic and metabolomic datasets. The system architecture is centralized around the ObjectStore, a Java based custom object/relational mapping system optimized for read-only access to a PostgreSQL database hosting the collated biological data. The query vocabulary is based around standard set theory and Boolean operators. Queries are interpreted by the Object-Store into SQL and results returned from the PostgreSQL database. The use of pre-computed tables of results, comprising a smart caching system, significantly enhances retrieval times, despite the significant (100s GB) amounts of data stored locally. Queries can be mediated through RESTful web services (currently capable for Python, Perl, Ruby, Java and Javascript) and directly via a web GUI. One particularly attractive aspect of the InterMine system is its innate interoperability - as the schema is based around Sequence Ontology, and because all living organisms possess genomes, it is possible to query across different InterMine instances for common homologues, enabling true comparative analysis [5]. Currently some 29 different instances [6] of InterMine are operating in different institutions, including a variant developed by us and colleagues at the Albert Einstein College of Medicine called ToxoMine [7] designed to host diverse datasets on the water-borne parasite Toxoplasma gondii with the goal of understanding virulence mechanisms associated with development of the disease toxoplasmosis. Here we describe recent developments to enhance the interoperability of the InterMine science gateway ecosystem, and pathways to optimize more effective use of InterMine - currently these systems require a level of bioinformatics expertise that is not common among the vast majority of the user community, especially by PIs who may wish to test specific hypotheses themselves. Bridging this gap would significantly enhance both wider user uptake and also overall scientific yield.

Keywords—data warehouse; systems biology; integrative analysis; InterMine; science gateways

REFERENCES

[1] A. Conesa, A. Mortazavi, "The common ground of genomics and systems biology," BMC Systems Biology, 8 Suppl 2:S1, 2014.

[2] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," Nature Genetics, 25, 1:25-9, 2000.

[3] A. Kalderimis, R. Lyne, D. Butano, S. Contrino, M. Lyne, J. Heimbach, F. Hu, R. Smith, R. Stepan, J. Sullivan, G. Micklem, "InterMine: extensive web services for modern biology," Nucleic Acids Res., 42, Web Server issue:W468-72, 2014.

[4] K. Eilbeck, S.E. Lewis, C.J. Mungall, M. Yandell, L. Stein, R. Durbin, M. Ashburner, "The Sequence Ontology: a tool for the unification of genome annotations," Genome Biology, 6, 5:R44, 2005.

[5] J. Sullivan, K. Karra, S.A Moxon, A. Vallejos, H. Motenko, J.D. Wong, J. Aleksic, R. Balakrishnan, G. Binkley, T. Harris, B. Hitz, P. Jayaraman, R. Lyne, S. Neuhauser, C. Pich, R.N. Smith, Q. Trinh, J.M. Cherry, J. Richardson, L. Stein, S. Twigger, M. Westerfield, E. Worthey, G. Micklem, "InterMOD: integrated data and tools for the unification of model organism research," Scientific Reports, 3:1802, 2013.

[6] http://intermine.org/

[7] D.B. Rhee, M.M. Croken, K.R. Shieh, J. Sullivan, G. Micklem, K. Kim, A. Golden, "toxoMine: an integrated omics data warehouse for Toxoplasma gondii systems biology research," Database (Oxford), 2015:bav066, 2015.