2: Biological Data Mining

Difficulties in data handling in the field of biology

If you are a researcher or deal with researchers as a part of your job, you probably know how difficult it is to handle data. While this statement is true in general, things get even worse when large databases from a range of heterogenous sources need to be brought together to allow further work and safekeeping. In the field of biology, using diverse datasets is becoming increasingly common, but it is very difficult in practice. This type of work, called integrative analysis, is possibly only when biological data is properly organised and can be queried easily.

The InterMine platform

Conscious of the above issues, researchers from the Micklem Lab at the University of Cambridge created InterMine, an open source data warehouse system for the integration and analysis of complex biological data. The system, which is still in active development, is now used by a number of major model organism databases. InterMine provides parsers for integrating data from many common biological data sources and formats, and also lets users add their own data. One of the most interesting features of the InterMine system is that it allows querying and data mining, even in the case of very large databases (e.g., the modENCODE projects contains >300GB of data). Such a feature is essential for researchers, as it would be very difficult for them to achieve the same level of performance independently.

The InterMine system was built with the uses in mind and provides an easy-to-use web application with advanced analysis tool. This allows end-users to access and explore data without any programming knowledge, which means that InterMine bridges a very wide gap for researchers. In addition, InterMine takes full advantage of the most recent cloud technology, allowing users to start instances of the system on the Amazon Cloud.

Impact and reach

As of March 2017, Google Scholar reports a grand total of 574 citations for the main individual data warehouses powered by the InterMine system (InterMine, FlyMine, MitoMiner, modMine, TargetMine, and YeastMine). By powering these, the InterMine data warehouse system indirectly allows advancements in biology research, including, but not limited to, the study of model organisms, mammalian localisation evidence, phenotypes and diseases, and drug discovery. The information in the data warehouses powered by InterMine allows biologists to re-use data that has been previously captured by other studies and build on these research findings, saving time and resources. The true reach of the Intermine system is likely to be even greater than the citation figures suggest, as data is often used without citations (as is often the case for freely available web resources).

Title	2: Biological Data Mining
Subtitle	Difficulties in data handling in the field of biology
Abstract	Biologists often need to deal with heterogeneous data sources, which makes their work difficult and time-consuming. The InterMine system provides an easy-to-use data warehouse solution that biologists can exploit for their studies with little programming knowledge.
Keywords	biology; data warehouse; data mining
Research subject area	BIOLOGICAL SCIENCES
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	N/A
Facts and figures	574 citations
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2012
Organisations involved	University of Cambridge
Academic citations	Smith, R.N. et al. (2012). [InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics.](https://doi.org/10.1093/bioinformatics/bts577)
Links	http://intermine.readthedocs.io/en/latest/about/
	http://www.modencode.org/publications/about/index.shtml
	http://www.flymine.org/
	http://mitominer.mrc-mbu.cam.ac.uk/
	http://intermine.modencode.org/
	http://targetmine.mizuguchilab.org/
	http://yeastmine.yeastgenome.org/