National Digital Research Infrastructure Organization

White Paper: Open Database Support and Curation

Authors

Nicholas Provart – University of Toronto nicholas.provart@utoronto.ca

Gary Bader – University of Toronto gary.bader@utoronto.ca

Lincoln Stein – Ontario Institute for Cancer Research lincoln.stein@oicr.on.ca

Michael Brudno – Hospital for Sick Children brudo@cs.toronto.edu

Summary

Biological databases and database-driven web servers are critical research infrastructure. Thousands of such resources have been developed. These are typically very difficult to maintain after very large initial investment to create them. Ideally, these servers would be deployed in a container on a public cloud, would automatically update themselves and be versioned, and be downloadable for anyone to install and modify locally to support sustainable development.

Background

With the publication of the first eukaryotic genomes around the start of the millennium, the past two decades have seen the advent and subsequent explosion of the field of bioinformatics, which encompasses the development and application of computational tools in managing all kinds of biological data. It involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, proteins and metabolites.

One important aspect of bioinformatics is the ability to access publicly-generated data in online databases. For instance, GenBank, the primary repository of nucleotide sequence information, permits the easy access and retrieval of nucleotide sequences (and computationally-derived protein sequences) at the click of a mouse. GenBank is run by the National Center for Biotechnology Information (NCBI) in the U.S., which is part of the National Library of Medicine (NLM), which in turn is part of the National Institutes for Health (NIH). The NCBI was established in 1988 after three independent but linked actions (see https://www.ncbi.nlm.nih.gov/books/NBK148949/ for a brief history): between 1984-86 advocacy groups held meetings on Capitol Hill to inform legislators and their staff on the value of the coming field of genomic research; in 1986 the NLM's Long Range Plan was finalized and it contained a recommendation that a new NLM Division be created to curate, manage and distribute molecular biology information; and in 1987 the House Select Committee on Aging, chaired by Senator Claude Pepper, introduced a bill to establish the NCBI to deal "with nothing less than the mystery of human life and the unfolding scroll of knowledge, seeking to penetrate that mystery, which is life itself." NCBI was funded by this bill initially at a level of $8 million per year, which supported a dozen staff members. Currently, the NCBI has created and maintains over 40 integrated databases for the medical and scientific communities as well as the general public. It received 3 million visitors daily (!) to its website and approximately 27 terabytes of data are downloaded per day.

Each year the January issue of Nucleic Acids Research focusses on new biological databases. There is now a plethora of such databases. Some are "consolidation tier" databases and serve an aggregating function for (model) organism-specific information (IAIC and others, 2010), such as The Arabidopsis Information Resource (TAIR) at arabidopsis.org, Wormbase for *C. elegans* researchers at wormbase.org, the Saccharomyces Genome Database at yeastgenome.org, etc. Other databases are domain-specific and curate data across different species for specific aspects of biology, like protein-protein interactions (e.g. BioGRID at thebiogrid.org; Chatr-Aryamontri et al., 2017), protein motifs (e.g. Pfam at pfam.xfam.org; Bateman et al., 2002) or protein structures (e.g. PDB at https://www.rcsb.org/).

How these databases are funded and maintained is eclectic. Some are funded by NIH grants (e.g. BioGRID receives funding through an NIH R01 grant, https://grantome.com/grant/NIH/R01-OD010929-14), while others operate as a freemium model: The Arabidopsis Information Resource, an important site for plant researchers, lost its NSF funding in 2009 and re-established itself as a subscription-based service, with university libraries as its primary subscribers, along with some plant biotechnology companies who pay higher subscription fees. Non-subscribers may access a limited number of TAIR pages per month before being blocked. Several other models for funding databases exist (Chandras et al., 2009) but most have limitations. For instance, advertising in the form of banner ads can seldom provide enough funds to pay for a curator. Some online resources, like GeneCards.org that consolidates information about human genes, include links to biotechnology reagents and charge $1-$4 per clicked reagent link (based on personal communication with Yaron Guan-Golan of LifeMap Sciences, which operates GeneCards.org). A further wrinkle to charging for access is presented by granting agencies' demands for data generated by public funds to be deposited in open repositories. Would a database that operates under a freemium model be considered open access?

One of the authors of this white paper, Nicholas Provart, runs the Bio-Analytic Resource for Plant Biology at bar.utoronto.ca (originally published as Toufighi et al., 2005 but now encompassing dozens of publications), which receives about 4 million page views per month, on par with TAIR. The BAR has one dedicated full time bioinformatician, who has been paid for by successive grants to Dr. Provart from Genome Canada to develop new software for these grants. But a good chunk of the bioinformatician's efforts are aimed at keeping the BAR in an operational state by updating packages and databases. Some aspect of plant protein-protein interaction curation has been done by undergraduate project students or casual employees. This really isn't a sustainable model, although it does provide good opportunities for gaining experience.

The GeneMANIA gene function prediction server (Warde-Farley et al., 2010) at www.genemania.org has a similar story to the Bio-Analytic Resource for Plant Biology. It was originally funded by a $2M grant from Genome Canada, followed by a 4-year grant from Ontario. Since then, maintenance falls on the Bader lab, who makes sure to regularly run the automatic build system and keep the servers functioning. GeneMANIA serves about 10,000 users per month.

We propose an allocation within any funding pool being proposed by the NDRIO towards supporting curation, databases and database-driven websites. The funds could be allocated through a competition, in 5-year blocks. Such an allocation would go a long way towards supporting and encouraging important online biological databases in Canada.

References

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L.L. (2002). The Pfam Protein Families Database. Nucleic Acids Res. 30: 276–280.

Chandras, C., Weaver, T., Zouberakis, M., Smedley, D., Schughart, K., Rosenthal, N., Hancock, J.M., Kollias, G., Schofield, P.N., and Aidinis, V. (2009). Models for financial sustainability of biological databases and resources. Database 2009.

Chatr-Aryamontri, A. et al. (2017). The BioGRID interaction database: 2017 update. Nucleic Acids Res. 45: D369–D379.

IAIC and others (2010). An international bioinformatics infrastructure to underpin the Arabidopsis community. Plant Cell 22: 2530–2536.

Toufighi, K., Brady, S.M., Austin, R., Ly, E., and Provart, N.J. (2005). The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. Plant J. 43: 153–163.

Warde-Farley, D. et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 38: W214-220.