REVIEW / SYNTHÈSE

Cheers for CHLA's 2009 bioinformatics course

Kathy Hysen

It was with trepidation and excitement that I enrolled in the bioinformatics course sponsored by the Canadian Health Libraries Association (CHLA) and taught by McGill University's Joan Bartlett, Assistant Professor in their School of Information Studies. Since I'm not currently working in a library environment, and with my nursing and science education a little rusty, I was looking for a topic that would be a fascinating "Entrez" to current scientific and information science practices. And fascinating it was…

As major breakthroughs continue in genetics research, its basics are becoming an integral part of elementary, secondary, college, and university curricula. These days, librarians do not usually require in depth knowledge about genetics to answer most questions; however, we need to consider the view 5 years from now. The human genome has already been sequenced, and the genetic basis for many human diseases will soon be mapped. Like the Millennials, who grew up being wired and connected, our clients will be familiar with aspects of genetics that still faze and amaze us. Increasing our subject knowledge can only improve our ability to meet the demands of our reference and instruction roles.

With this in mind, 26 science librarians from universities across Canada, one new Faculty of Information Studies graduate, and I arrived at the University of Toronto's Gerstein Science Information Centre early in the morning on June 8 to begin our investigations. With heartfelt thanks to the wonderful organizational skills of Ilo-Katryn Maimets and Gail Nichol, we were pampered with technology that functioned and delicious refreshments to sustain our intellectual efforts.

For the next amazing and intense 3 days, Professor Joan Bartlett energetically whirled and twirled our little grey cells through the intricacies of basic genetics, types of resources, typical reference questions, and some tips on how to maintain our equilibrium while navigating the wealth of resources available in this field. Last but not least, Professor Bartlett articulated some of the challenges associated with the design, usability, and authority of these databanks.

We focused on the three types of resources that would be accessed for questions that lie beyond the scope of basic molecular biology reference sources. These online collections consist of summaries of research illustrating the genetic connections with human diseases (e.g., OMIM (http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim)); databases of genetic sequences (e.g., UniProt (http://www.uniprot.org)); and tools that analyze these sequences for similarities (e.g., ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html)).

We are all familiar with the dynamic nature of electronic resources; their user interfaces, content, and ownership are not static by any stretch of the imagination. In genetics, this issue is compounded by the individuality of scientists and their research, their funding situations, and their particular computer programming colleagues. With the fast pace of recent genetic research, staggering numbers of new in-house databases have been built, with all their attendant idiosyncracies. An examination of the 2009 *Nucleic Acids Research* annual database issue (available at http://nar.oxfordjournals.org/content/vol37/suppl_1/index.dtl) drives home this point, with a total of 179 databases described, including 95 new ones.

Combined with the pivotal role that computer science plays in the structure of these resources, end-user issues for quality data mining have often been glossed over, if not ignored. Yet analysis and visualization of this data are the raisons d'être for these collections, since the days of just creating repositories of data are over. For example, the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/projects/geo/), European Bioinformatics Institute (EBI) ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/), and the University Health Network's MicroArray Centre (http://data.microarrays.ca/) here in Ontario contain similar data, with overlapping content. However, each one offers a vast number of ways to search for and examine their data.

Terminology is another aspect that has become quite complex due to the rapid growth of the field. Terms may or may not be referenced to synonyms, MeSH, or UMLS subject headings. Naturally, this leads to difficulties in accessing the most appropriate resource, using it efficiently, and accurately evaluating that resource. Gene Ontology (GO; http://www.geneontology.org/) is an international collaboration of sequence database developers that has led to the establishment of structured vocabularies. However, without an international agreement on bibliographic control for these types of resources, usability problems will only worsen.

Finally, content is not always authoritative (i.e., curated) nor transparent. Submissions are often accepted as is, in order to collect as much data as possible. For instance, major players such as NCBI clearly distinguish between GenBank (http://www.ncbi.nlm.nih.gov/Genbank/), their "as is" repos-

K. Hysen. 34 Askin Street, London, ON N6C 1E3, Canada (e-mail: kathy.hysen@sympatico.ca).

itory of genetic sequences, and RefSeq, their collection of GenBank records curated by subject experts. However, some collections are computer curated (e.g., TrEMBL records before being included in the Swiss-Prot database (http://www.ebi.ac.uk/uniprot/Documentation/index.html#TrEMBL)). There is a lot of fine print to read when investigating these complex resources.

I would highly recommend this bioinformatics course to other science librarians. However, I would also like to see it develop into something beyond a survey course so that more time is available for hands-on experience. One possibility would be to have an overview of one type of resource in the mornings, with practice in the afternoons. Perhaps the classroom sessions could be augmented by online tutorial practice questions, available as part of the continuing education hours of the course. Another possibility would be to have this course as the first in a series that becomes increasingly oriented toward resolving bibliographic issues of search and retrieval, as well as developing standards for authority control, classification, and controlled vocabulary.

Until then, perhaps we could collect questions and answers or notes through a CHLA blog, wiki, or Twitter, since these databases are numerous and can be quite dynamic and ephemeral. Perhaps some of us could form a core group to investigate and describe bioinformatics resources more rigorously. That core group could go on to communicate with scientists and associations in Canada and internationally about storage and retrieval questions. And perhaps we could make a difference.