High performance computing and medical research

A. Jamie Cuticchia

Bioinformatics, the use of computer technology to solve biological and medical problems, is a relatively new discipline.[1] Toronto's Hospital for Sick Children established a centre for bioinformatics (see www.bioinformatics-canada.org/) in 1998. The motivation to invest in this infrastructure was to provide researchers with a competitive advantage in coping with massive amounts of scientific data. The Bioinformatics Supercomputing Centre is the 455th-most powerful computing centre in the world and the largest centre devoted solely to biomedical research. Its computing power is equivalent to 1700 personal computers.

The need for this level of computing resource is driven both by the amount and complexity of biological data; more biological data will be generated in the year 2000 than the total of all data collected up to that point. The medical researcher is, therefore, burdened with the prospect of sifting through this information to cull the data that is relevant to a particular query. Although the ability to quickly sort biological data is aided by supercomputers that are configured to maximize efficiency at data retrieval or complex analysis, for scientists to use the biomedical software effectively, they must also must be trained in how to conduct different types of analyses properly. The Bioinformatics Supercomputing Centre provides researchers with training on the principles and tools of basic bioinformatics analyses (e.g., DNA sequence-homology searching,[2] protein folding,[3] gene-finding algorithms[4]), as well as help using the software. However, the researcher with a DNA sequence of interest can literally spend weeks performing complex data analysis, even on the fastest supercomputers. The implementation of a successful data acquisition, curation and dissemination system[5] and the modelling of biological data are key components of many scientific endeavors, the most relevant of which is the human genome project.[6]

Computational resources

The Origin 2000 serves as the workhorse for the Bioinformatics Supercomputing Centre. The computational power of the Origin 2000 is supplemented by an IBM RS/6000 SP3 system, a group of 5 nodes that are accessed independently and optimized for database use. In addition to providing a suite of programming languages, the centre hosts several bioinformatics applications. The most widely used application, BLAST (Basic Local Alignment and Search Tool), is accessible through a Web interface developed at the centre (http://blast.bioinfo.sickkids.on.ca/index.html); this application is designed to compare protein and nucleic acid sequences against a selection of genetic databases to calculate percent homology.

The system also supports the Genome Database (www.gdb.org), the official central repository for genomic mapping data resulting from the Human Genome Initiative; this initiative is a worldwide research effort to analyze the structure of human DNA and determine the location and sequence of the estimated 100 000 human genes. In support of this project, the Genome Database stores and curates data that is generated worldwide by researchers engaged in the mapping effort of the Human Genome Project;[8] it is a repository of human maps and map objects. The

maps include those produced by contig mapping, radiation hybrids, linkage studies and integrated data. The objects in the database are genes, amplimers, probes, sequence-tagged sites, polymorphisms and mutations. All data are peer reviewed by a group of approximately 100 editors. Links are also maintained to other databases, most notably to GenBank,[9] a database of nucleotide sequences from more than 58 000 organisms.

The Genome Database is accessed over 15 million times a year at the central node in Toronto alone. Additionally, 12 mirror sites provide a current copy of the database to users. Remote nodes were established not only to provide quicker access time for users but also to provide points where those using or submitting data to the database could receive support from other researchers in their own time zone and in their native language. These sites can provide both researchers and the general public with information on genetics. For example, by querying with simple text such as "cystic fibrosis," one can retrieve the most relevant citations in the area, a list of the markers for the gene, a list of mutations, as well as links to clinical information.

Although Canada does not presently have an institute devoted to bioinformatics, the diversity in the ongoing research here is likely to ensure that Canada will be an international force in this field. Two conditions in Canada presently endanger the development of bioinformatics. First, there is the fear that bioinformaticians might be seen as a "homogeneous" group of scientists and that the coordination of efforts might be imposed on various research groups and thus stifle creativity. Second, there exists no official government funding mechanism to support bioinformatics research. As universities across Canada recruit new bioinformaticians, program funding and resources should be devoted to fuel the discipline's growth.

Dr. Cuticchia is Head of Bioinformatics and Senior Bioinformatics Scientist, Research Institute, The Hospital for Sick Children, Toronto, Ont.

Competing interests: None declared.

References

1. Boguski MS. Bioinformatics. *Curr Opin Genet Dev* 1994;4(3):383-8.
2. Altshul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215(3):403-10.
3. Sternberg MJ, Bates PA, Kelley LA, MacCallum RM. Progress in protein structure prediction: assessment of CASP3. *Curr Opin Struct Biol* 1999;9(3): 368-73.
4. Uberbacher EC, Xu Y, Mural RJ. Discovering and understanding genes in human DNA sequence using GRAIL. *Methods Enzymol* 1996;266:259-81.
5. Cuticchia AJ, Chipperfield MA, Porter CJ, Kearns W, Pearson PL. Managing those bytes: the human genome project. *Science* 1993;262:47-8.
6. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gasteland R, Walters L. New goals for U.S. Human Genome Project: 1998–2003. *Science* 1998;282: 682-9.
7. Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, et al. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* 1989;245(4922):1059-65.
8. Talbot CC Jr, Cuticchia AJ. Human mapping databases. In: Dracopoli NC, Haines J, Korf BR, Morton C, Seidman CE, Seidman JG, editors. *Current protocols in human genetics*. New York: John Wiley and Sons; 1999. p. 1.13.1-12.
9. Benson DA, Boguski MS, Lipman DJ, Ostell J, Oullette BF, Rapp BA, et al. GenBank. *Nucleic Acids Res* 1999;27(1):12-7.

Correspondence to: *Dr. A. Jamie Cuticchia, Bioinformatics Supercomputing Centre, The Hospital for Sick Children, 555 University Ave., Toronto ON M5G 1X8; fax 416 813-8755.*