Keeping pace with the data: 2008 update on the Bioinformatics Links Directory

Michelle D. Brazas[1], Joanne A. Fox[2], Timothy Brown[1], Scott McMillan[3] and B. F. Francis Ouellette[1],*

[1]Ontario Institute for Cancer Research, 101 College St, Suite 800, Toronto, Ontario, Canada M5G 0A3, [2]University of British Columbia, Michael Smith Laboratories and [3]University of British Columbia, Office of Learning and Technology, Vancouver, British Columbia, Canada

ABSTRACT

The Bioinformatics Links Directory, http://bioinformatics.ca/links_directory/, is an online resource for public access to all of the life science research web servers published in this and previous issues of *Nucleic Acids Research*, together with other useful tools, databases and resources for bioinformatics and molecular biology research. Dependent on community input and development, the Bioinformatics Links Directory exemplifies an open access research tool and resource. The 2008 update includes the 94 web servers featured in the July 2008 Web Server issue of *Nucleic Acids Research*, bringing the total number of servers listed in the Bioinformatics Links Directory to over 1200 links. A complete list of all links listed in this *Nucleic Acids Research* 2008 Web Server issue can be accessed online at http://bioinfomatics.ca/links_directory/narweb2008/. The 2008 update of the Bioinformatics Links Directory, which includes the Web Server list and summaries, is also available online at the *Nucleic Acids Research* website, http://nar.oxford journals.org/.

COMMENTARY

The past several years have seen the introduction of several new and advanced experimental technologies in the biological sciences. These technologies, which include next generation sequencing and imaging as well as various other nanoscale experimental processes, have dramatically increased the throughput capacity of life science research, and have also been the source for an unprecedented volume of experimental data. Given the quantity and variety of data being produced, research scientists can now ask more probing biological questions to gain insight on such curiosities as the interactions, pathways and networks at play in a given disease or biological function, or ask questions that explore the commonalities and variations between large data sets from different macromolecules, species or organisms.

Keeping pace with these advances in technology and data output has been the number of specialized web servers and bioinformatic resources developed or upgraded to meet these new data intensive research needs. Since 2004, *Nucleic Acids Research* has peer-reviewed and published in their Web Server issue, a compendium of the latest web servers and freely available online bioinformatic tools to keep researchers abreast of the deluge of bioinformatic resources available to them. This year's Web Server issue introduces an additional 94 bioinformatics and molecular biology web servers, 10 of which are updates (Table 1). Along with the long-standing Database issue (1), the special Web Server issues represent an invaluable source of bioinformatic tools and resources for the international life-science research community. The complete listing of URLs cited in the 2008 Web Server issue can be accessed online at the *Nucleic Acids Research* website, http://nar.oxfordjournals.org/, as well as at http://bioinfomatics.ca/links_directory/narweb2008/.

The Bioinformatics Links Directory, http://bioinformatics.ca/links_directory/, is a public, curated collection of all of these servers together with other useful tools, databases and general purpose resources for bioinformatics and molecular biology research. Since 2005, *Nucleic Acids Research* has partnered with the Bioinformatics Links Directory to ensure that all of the links published in the Web Server special issues are included in the directory (2–4). This 2008 update brings the total number of servers and tools listed in the Bioinformatics Links Directory to over 1200 unique links (Table 1).

Organized by biological subject with subcategories of common tasks relevant to the subject, the Directory serves as a 'go-to' site for the research community seeking bioinformatic resource options. Each entry contains a short description of the tool's function as well as the accompanying PubMed citation and web server URL.

*To whom correspondence should be addressed. Tel: +1 416 673 8511; Fax: +1 416 977 7446; Email: francis@oicr.on.ca

Table 1. Summary of the number of web servers listed in each sub-category of the Bioinformatics Links Directory

Name	URL[a]
Computer Related	
Bio-* Programming Tools	20
C/C++	3
Databases	2
Java	4
Linux/Unix	11
PERL	5
PHP	1
Statistics	9
Web Development	2
Web Services	7
DNA	
Annotations	57
Gene Prediction	34
Mapping and Assembly	15
Phylogeny Reconstruction	46
Sequence Feature Detection	145
Sequence Polymorphisms	41
Sequence Retrieval and Submission	32
Tools For the Bench	65
Utilities	23
Education	
Bioinformatics Related News Sources	9
Community	23
Courses, Programs and Workshops	5
Directories and Portals	15
General	14
Tutorials and Directed Learning Resources	9
Expression	
cDNA, EST, SAGE	44
Gene Regulation	120
Microarrays	101
Protein Expression	17
Splicing	19
Networks	8
Human Genome	
Annotations	38
Ethics	8
Genomics	3
Health and Disease	23
Other Resources	29
Sequence Polymorphisms	36
Literature	
Goldmines	6
Open Access Resources	2
Search Tools	12
Text Mining	22
Model Organisms	
Fish	11
Fly	17
General Resources	28
Microbes	45
Mouse and Rat	35
Other Organisms	21
Other Vertebrates	10
Plants	21
Worm	9
Yeast	18
Other Molecules	
Carbohydrates	6
Metabolites	4
Small Molecules	6
Compounds	2

(continued)

Table 1. Continued

Name	URL[a]
Protein	
2-D Structure Prediction	60
3-D Structural Features	75
3-D Structure Comparison	50
3-D Structure Prediction	60
3-D Structure Retrieval, Viewing	52
Biochemical Features	41
Do-it-all Tools for Protein	13
Domains and Motifs	115
Function	47
Interactions, Pathways, Enzymes	94
Localization and Targeting	38
Molecular Dynamics and Docking	27
Phylogeny Reconstruction	45
Presentation and Format	14
Protein Expression	8
Proteomics	33
Sequence Data	9
Sequence Comparison	7
Sequence Features	33
Sequence Retrieval	29
RNA	
Functional RNAs	26
General Resources	10
Motifs	22
Sequence Retrieval	11
Structure Prediction, Visualization, and Design	54
Sequence Comparison	
Alignment Editing and Visualization	21
Analysis of Aligned Sequences	60
Comparative Genomics	35
Multiple Sequence Alignments	56
Other Alignment Tools	11
Pairwise Sequence Alignments	26
Similarity Searching	47

[a]A complete listing of all URLs listed in the *Nucleic Acids Research* 2008 Web Server Issue can be accessed online at: http://bioinformatics.ca/links_directory/narweb2008/

The subject categories and subcategories are easily browsed and queried with a keyword search. Among the new web resources for 2008, are those listed under 'Networks', a new subcategory under 'Expression' (Table 1), representing the need and introduction of new resources for the integration of expression data from various studies.

The Bioinformatics Links Directory is also an excellent example of a community resource driven by researchers who consider free and public access to their work essential to the progress of science. Suggestions for new links or revisions and corrections to existing links at the Bioinformatics Links Directory are welcome, and may be submitted through email directly to links@bioinformatics.ca. The up-to-date complete listings accessible through the Bioinformatics Links Directory, including the *Nucleic Acids Research* 2008 web servers, is available online at http://bioinfomatics.ca/links_directory/narweb2008/.

In looking forward as research technologies and platforms continue to advance, the web will continue to play an increasing role as a data source. Already, the web has become an important mechanism for the communication, access and exchange of data. As noted by Fox *et al.* (4),

blogs, application programming interfaces (APIs), wikis and really simple syndication (RSS) feeds are extending the communication capacity and information output of the web. However, with the current pace of data output and the increasing need to synthesize research data from multiple sources, even use of the web to identify, access and extract meaningful information for research purposes is becoming a daunting task. This exponential explosion of information in science, compounded by the specialization and heterogeneity of the information, simply overwhelms any one individual's ability to store and model all of the relevant science in their head (http://sciencecommons.org/projects/data).

However, changes in how the web's content is organized and structured offer the opportunity to automate computers to navigate and integrate all of the biological information stored on the web, and output coalesced information to the researcher for interpretation. The Semantic Web is an extension of the current web and is based on common formats that enable automated navigation and integration of data from diverse sources (5,6). Rather than the web being a decentralized platform for the distribution of 'presentations' of information, the semantic web is a decentralized platform for the distribution of 'knowledge' (http://www.w3.org/2001/sw/), which can be shared and used across applications and research community boundaries because the format of semantic web data allows for data integration if the sources describe the same biological entity. For example, current links on web pages are uncharacterized so that there is no explicit information to tell a computer that the Bioinformatics Links Directory for the BLAST tool (7) that finds regions of local similarity between sequences, is in any way related to another directory entry for the T-Coffee tool (8) for protein multiple sequence alignment. However, in the Semantic Web, because relationships are captured in 'subject-relationship-object' statements using Uniform Resource Identifiers (URIs) (http://www.rfc-editor.org/rfc/rfc3986.txt), the relationship between the BLAST and T-Coffee tools can be readily identified by a computer. Whenever two subjects (in this case BLAST and T-Coffee) refer to identical URIs (in this case capacity for protein sequence alignment), then their topics of discourse are identical and data merging becomes possible. The Semantic Web is thus a means to capture and network the relationships implicit in high volume data sets, or in the outputs of sophisticated analytic software, because anything can be related to anything, as long as that anything has a unique name or URI (http://sciencecommons.org/projects/data). Applications of the Semantic Web are being explored in neuroscience (6) (http://www.w3.org/2001/sw/hcls/) with some impressive and promising results for the future of biological research on the web.

Using the Semantic Web, researchers will thus be able to input a gene of interest from an experiment into a computer and explicitly ask the computer to return information on how this gene functions in another organism, or how the product of this gene affects a given biological process, or which compounds also affect that biological process and whether these compounds have been shown to have the same affect in other organisms. The current structure of the Bioinformatics Links Directory is amenable to semantic web notation and upgrading of the directory to encompass this functionality is being explored. While adoption of the semantic web into biological research is not without its challenges, the potential power, knowledge and discoveries to be gained from integrating and networking the already complex and diverse biological data, should be a sufficient driving force for exploiting the web in today's research arena.

REFERENCES

1. Galperin,M.Y. (2008) The Molecular Biology Database Collection: 2008 update. *Nucleic Acids Res.*, **36**, D2–D4.
2. Fox,J.A., Butland,S.L., McMillan,S., Campbell,G. and Ouellette,B.F. (2005) The Bioinformatics Links Directory: a compilation of molecular biology web servers. *Nucleic Acids Res.*, **33**, W3–W24.
3. Fox,J.A., McMillan,S. and Ouellette,B.F. (2006) A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Res.*, **34**, W3–W5.
4. Fox,J.A., McMillan,S. and Ouellette,B.F. (2007) Conducting research on the web: 2007 update for the Bioinformatics Links Directory. *Nucleic Acids Res.*, **35**, W3–W5.
5. Berners-Lee,T. and Hendler,J. (2001) Publishing on the semantic web. *Nature*, **410**, 1023–1024.
6. Ruttenberg,A., Clark,T., Bug,W., Samwald,M., Bodenreider,O., Chen,H., Doherty,D., Forsberg,K., Gao,Y., Kashyap,V. *et al.* (2007) Advancing translational research with the Semantic Web. *BMC Bioinform.*, **8 (Suppl 3)**, S2.
7. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
8. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.