The biologist and the World Wide Web: an overview of the search engines technology, current status and future perspectives

Hervé Récipon* and Wojciech Makalowski†

Addresses
National Center for Biotechnology Information, National Institutes of Health, Building 38A, Room 8N-805, Bethesda, MD 20894, USA
* e-mail: recipon@ncbi.nlm.nih.gov
† e-mail: makalow@ncbi.nlm.nih.gov

Current Opinion in Biotechnology 1997, **8**:115–118

Electronic identifier: 0958-1669-008-00115

Abbreviations

CGI	Common Gateway Interface
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
Web	World Wide Web
WWW	World Wide Web

Introduction

In 1990, very few people were able to predict the impact of the Internet (see [E1] for general information about the Internet) and the World Wide Web (WWW or Web) on molecular biology research and in particular on the genome projects which were in their infancies. After only six years, thousands of Web sites pertaining to biology have been created worldwide. Accessed thousands of times per day, these Web resources are now a part of a scientist's daily 'routine'. They provide vital information, allow users to exchange views and ideas, and help biologists in the analyses of their data. By 'cross-referencing' ('linking' for short) to others sites, the webmasters (title frequently used to describe persons performing the many duties involved in creating and maintaining a Web resource) have created a valuable network of tools considered indispensable for modern research.

Another important benefit for researchers around the world is that any data published in a web site is available virtually instantaneously. As a result, this has changed the way scientists reports their results. Formerly, the goal of publishing was to make results available via scientific journals found only in the laboratories' libraries. Access to that type of scientific information was, therefore, restricted to a limited audience. Today, in the 'Web world', the new goal is to rapidly make the information 'available on the internet' ('online' for short) via local institutional Web sites or via a scientific journal web site (a majority of journals are currently available online; see [E2] for the most current complete list). Using the Web efficiently facilitates the publication of scientific results in a 'human' readable format; consequently, in the Web world, access to scientific news and information is now quickly available to almost everyone. As shown by a few of the recent discovery announcements, the press conferences and the availability of the information on the Web are now often synchronized to maximize the impact of the announcement. The scientific Web sites have therefore become the research showrooms for the general public, and they are an integral part of the media that are omnipresent today.

The origin of this revolution is not in the programs or resources themselves; these programs and resources, which are usually complex and expensive to maintain, were reserved, just only a few years ago, for a limited group of experts in well-budgeted institutions. Now they are available to everyone using a simple cheap personal computer connected to the Internet. This is not just due to a simple change in the molecular biology technology alone (which has also evolved greatly in the past six years): this is a revolution in the accessibility of the technology that changes the way people are thinking, conducting and publishing their research.

In this paper, we present a overview of current and future Web technology, and the terminology of the Web in order to provide enough information to allow the user to better understand how a Web server is working. A full description of the available Web technology is beyond the scope of this discussion, but more detailed information can be easily found in a number of useful guidebooks [1,E3].

The Web interface

In practice, the Web is a vast collection of interconnected documents, spanning the world and accessible by using their Uniform Resource Locator (URL). The documents displayed by the Web browser are special hypertext documents that are using the Hyper Text Markup Language (HTML) [E4]. The advantage of hypertext is that in a hypertext document, if you want more information about a particular subject mentioned, you can usually 'just click on it' to read about it in further detail. In fact, HTML documents can be and often are linked to other documents by completely different authors — much like footnoting, but the referenced document is retrieved instantly! The Web browsers allow the user to use the links in a transparent way: select the link and you are presented with the text that is pointed to. In fact, today, the current HTML version allows Web browsers to display and 'point' to a text file, an image, a sound, an animation, or a mix of all four. In a multimedia World Wide Web, any medium can point to any other medium.

The HTML embeds special tags that describe the structure as well as the look of a document or section of text. Web browsers translate HTML, which is basically an ASCII text, into the multimedia documents. Web browsers interpret the HTML commands and display them on the computer screen. As most Web browsers are written for many different computer platforms with versions for PCs, Macintoshes, UNIX and other systems, this allows the same Web page to look the same on every computer.

There are two kinds of HTML documents: static and dynamic. Static documents are delivered from the Web server using the Hyper Text Transfer Protocol (HTTP). The HTTP server (HTTPd) reads a file located on the hard drive of the Web server and sends it to the Web browser. Static documents are usually used to provide information that does not change or can be easily modify by hand. In contrast, dynamic documents are generated 'on the fly' by a program using the Common Gateway Interface (CGI) [E5]. The dynamic documents usually never really exist on the hard drive of the Web server. Each type has its own pros and cons. Static documents are easy to create, especially with the latest easy-to-use WYSIWYG (What You See Is What You Get) editors. They are usually delivered to the Web browser faster than the dynamic documents but they cannot accept input from the user and, worst of all, they have a high maintenance cost. Alternatively, dynamic documents can be created to have access to any kind of database (such as Genbank), from any scientific instruments (such as weather monitoring systems), or to any scientific programs (such as a sequence similarity search program like BLAST).

These CGI programs (or CGI-bin) usually not only create the dynamic document that allows the user to submit their request via a FORM (one of the HTML tags) but also to interpret this request from the Web browser, interact with the database/instrument/program and generate the HTML document which is sent back to the Web browser. Because of its relative ease of use, this technology provides an elegant way of obtaining a stable client/server system with a high level of portability over the different computer platforms. The directness of the CGI-bin interface simplifies the work of the webmaster and allows him or her to concentrate his or her effort only on the service he or she wants to provide. In addition to this, the current state of the art in CGI-bin technology is also to control the FORM created by the program with a JavaScript to detect early obvious errors. JavaScripts are also used to allow the user to interact with the HTML document.

What is JavaScript? JavaScript [E6] is a compact, object-based scripting language for developing small Web applications and it is currently only used in Netscape Navigator (version 2.0 and higher) [E7]. Netscape Navigator interprets JavaScript statements embedded directly in an HTML page. At the level of the Web browser, those statements can recognize and respond to user events such as mouse clicks, form input and page navigation. For example, you can write a JavaScript function to verify that users enter valid information into a FORM requesting a GenBank sequence entry. Without any network transmission, an HTML page with JavaScript embedded can interpret the entered text and alert the user with a message dialog if the input is invalid. Or you can use JavaScript to perform an action in response to the user opening or exiting a page. This adds a second layer of complexity to the Web and also increases its capabilities. The Web browser client is not a merely a dull 'displayer' presenting the documents received from the Web server; with the addition of JavaScript, the browser now is also actively responsible for some of the Web features described above. This presents the obvious advantage of a quicker response, because the server is not involved in executing JavaScript.

The molecular biology tools available on the Web

The biological community was one of the first that took the advantage of Web technology for their research. The Internet is probably the most complete source of information accessible more or less in a single place because biologists can obtain all this information without moving from their laboratory. Moreover, more and more scientific journals (such as *Science*, *Genome Research* and *Gene*) have some additional appendices only available online. Usually, biological Web sites can provide three kind of information. The first two use static HTML documents whereas the third uses dynamic HTML. Web sites are used by biologist for posting all sorts of information; they consist of collections of links to specific pages or serve as links between users and programs that are running on remote computers. In fact, a majority of Web sites are mixtures of these categories.

Thanks to Web technology, huge databases and programs that require powerful computers became easily accessible for most biologists directly from their laboratories. Most computational biology problems (such as sequence analysis, multiple sequence alignments, database requests, sequence similarity searches and structural retrieval etc.) can now be solved by accessing the appropriate Web site. Below, we present some examples of software tools available for molecular biologists through the WWW interface. The more extensive lists of such tools are available in several places, including Pedro's BioMolecular Research Tools site [E8].

Access to both general and specialized databases is available on the Web. One the most interesting tools is the Entrez system at the National Center for Biotechnology Information in Bethesda, USA [E9]. Entrez is a sophisticated system that combines four different type of data: literature (MEDLINE), nucleotide (GenBank), protein

(SwissProt and PIR [Protein Information Resource]) and protein structures (PDB; Protein Data Base). Any database can be used as a starting point for data retrieval. For example, searching for 'aspartyl tRNA synthetase' leads one to two records in the protein structure database with four links to Medline, and two links to both nucleotide and protein databases. Further exploration of the information on the given subject is possible through the precomputed 'similar' records in all databases. The Entrez system is an example of a general database. A number of specialized databases are also available on the Web. They can be focused on certain biological problems, for example, BLOCKS at the Fred Hutchinson Cancer Research Institute is a database of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins [E10]. Other specialized databases may be devoted to single organisms like DictyDB — a database for *Dictyostelium discoideum* at the University of California at San Diego (UCSD) [E11].

Another group of biological services contains programs for the analysis of data. Tools found on the Web can be as simple as the translation of nucleotide sequences into proteins (e.g. the Translate program at the ExPASy server [E12]) or as sophisticated as gene prediction (e.g. Grail at Oak Ridge National Laboratory [E13]). Especially useful for molecular biologists are servers that offer a wide spectrum of tools for sequence analysis. One of the best such places is the ExPASy server at the University of Geneva, Switzerland [E14]. As the home of SwissProt, the place is focused on protein analysis. The server offers access to several databases such as the curated protein database (SWISS-PROT), the enzyme nomenclature database (ENZYME), the sequence analysis bibliographic data bank (SeqAnalRef), the two-dimensional polyacrylamide gel electrophoresis database (SWISS-2DPAGE) and others. The ExPASy offers a variety of tools for sequence analysis. The programs run directly on the ExPASy server or exist as links to remote sites. Tools offered here include protein identification, nucleotide sequence translation, similarity searches, pattern and profile searches, primary sequence analysis, secondary structure prediction, tertiary structure analysis, and sequence alignment. In addition, the ExPASy server contains some links to other molecular biology servers.

As described above, the Web is an excellent resource of biological data and tools. The WWW is growing so fast that finding sites of special interest is no trivial task any more. Web search engines like Alta Vista or Yahoo very often return too many hits. In this situation, Web sites that are collection of links to other sites of interest became very useful. Two such sites are especially popular among biologists: Keith Robison's World Wide Web Virtual Library [E15] and Pedro's BioMolecular Research Tools [E8]. In Europe, the Pasteur Institute offers a powerful server to search for WWW biological sites [E16].

Next step: the distributed objects and Java

If you are just starting to feel comfortable with the above and your current Internet applications, you may not like to read the following: the computer industry is undergoing a second client/server revolution. This revolution promises to be at least as traumatic as the one we went through when giant mainframe-based applications were broken apart into client and server components. This second revolution is named 'distributed objects' [2].

A classical computer object encapsulates code and data. These objects provide wonderful code-reuse facilities, but they live only in one program and on one computer at a time. The outside world doesn't know about these objects and, worse, has no way to access them. In contrast, a distributed object can live anywhere on the network and can be accessed by remote clients. Clients don't need to know where the distributed object resides or what operating system it executes on; it can be on the same machine or a machine that sits across the planetary network and it can be accessed by any kind of programming languages (C, C++, Smaltalk, Java etc.). To create this kind of intercomputer, inter-operating system, interlanguage exchangeable smart object, the computer industry needed to create a standard. In fact, there are two standards: OLE/COM (Object Linking and Embedding/Common Object Model) [E17] and OpenDoc/CORBA (Common Object Request Broker Architecture) [E18,E19]. Currently, CORBA is supported by more than 500 companies, making it the largest standard in existence, whereas OLE has been defined only by Microsoft.

The other component of the next generation of Web application is the programming language, which will use the CORBA standard, Java. Java is a language developed by Sun Microsystems that allows Web documents to contain a code that is executed on the browser. Because Java is based on a single 'virtual machine' that all implementations of Java emulate, it is possible for Java programs to run on any system that has a version of Java. It is also possible for the 'virtual machine' emulator to make sure that Java programs downloaded through the Web do not attempt to do unauthorized things (such as reading/writing a file or introducing a virus on your computer). Actually, Java can be used in the absence of the Web, but the application that has sparked so much interest in Java is HotJava, a Web browser written in the Java language. You can learn more about Java and HotJava from Sun's HotJava home page [E20,E21].

Confused by the names Java and JavaScript? Although they are related and JavaScript borrows most of Java's syntax, they are fundamentally different and serve different purposes. JavaScript is a simple object-oriented scripting language suitable for any small application exclusively on the Web whereas Java is a compiled complex complete object-oriented language suitable for consequent application not only restricted to the Web. They are complementary rather than competing with each other [E22].

Molecular biology tools using Java/CORBA on the Web

What benefits can computational biology expect from Java/CORBA? Most of the benefits are at the computer scientist (developer) level but they will also greatly affect biologists. Java/CORBA bring to the developer a high level of portability and reusability. Portability makes a program immediately available to all computer platforms. This satisfies the small developer who is quickly able to reach a wide range of users as well as the larger developer, because portability reduces the cost of program development on several operating systems. The reusability of the distributed objects that constitute a program also make life much easier for developers. They don't have to 'reinvent the wheel' for each program; instead, developers can build a new program using distributed objects like Lego® blocks, which allow them to concentrate on the final goal. The other benefit of this 'Legobuilding' is for users: if users do not need a full package, they will be able to only buy the parts they need. Finally, the use of the Web as client/server will create resources instantaneously accessible over the Internet. Java/CORBA means more creativity, easy development at lower cost, instant Internet accessibility and better fitting to the needs of the user.

Biologists will see an explosion of new applications on the Internet as well as on their local computer, with amazing new capabilities to change and adapt these programs to their needs themselves. One of the goals of the Bioinformatic Java/CORBA Working group [E23] and the bioViews Consortium [E24] is to regroup information about this new technology and to share the efforts of the computational biologists. These sites and the Java-based Molecular Biology Work Bench [E25] at the European Molecular Biology Laboratory (EMBL) also regroup the currently available demonstration programs. One of the immediate benefits of using Java/CORBA can be see in some of these demos. In particular, JADE [E26], the result of a collaboration between L Stein (Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, USA) and J Thierry-Mieg (Centre de Recherche en Biochimie Macromoleculaire-Centre National de la Recherche Scientifique [CRBM-CNRS], Montpellier, France), is a Java-based viewer for ACEDB (a *Caenorhabditis elegans* database) software to display genetic and physical maps of various genomes, as well as the PDB viewer [E27] from D Walther, EMBL, the Human Physical Map Viewer from Washington University [E28] and CINEMA (Colour INteractive Editor for Multiple Alignments) [E29] from D Parry-Smith, AWR Payne, AD Michie and TK Attwood. All these demos, although still at a preliminary stage, need Netscape version 2.0 or higher.

References

1. Stewart JM: *World Wide Web Encyclopedia CD.* Rockland, MA: Charles River Media Inc; 1996.

2. Orfali R, Harkey D, Edwards J: *The Essential of Distributed Objects.* New York: Willey & Sons Inc; 1996.

Electronic references

E1. 'About the Internet.' http://home.netscape.com/home/about-the-internet.html

E2. 'Bio/Chemical Journals and Newsletters.' http://www.public.iastate.edu/~pedro/rt_journals.html

E3. 'An Introduction to the World Wide Web.' http://cimesg1.epfl.ch/expose/plan.html

E4. 'NCSA Beginner's Guide to HTML.' http://www.ncsa.uiuc.edu/General/internet/WWW/HTMLPrimer.htm

E5. 'The Common Gateway Interface.' http://hoohoo.ncsa.uiuc.edu/cgi/

E6. 'Netscape Javascript.' http://home.netscape.com/comprod/products/navigator/version_2.0/script/

E7. 'Netscape Navigator 3.0.' ftp://ftp.netscape.com/pub/navigator/3.0

E8. 'Pedro's BioMolecular Research Tool.' http://www.public.iastate.edu/~pedro/research_tools.html

E9. 'National Center for Biotechnology Information.' http://www.ncbi.nlm.nih.gov/

E10. 'BLOCKS WWW Server.' http://www.blocks.fhcrc.org/

E11. '*Dictyostelium discoideum* Developmental Gene Program.' http://glamdring.ucsd.edu:80/others/dsmith/dictydb.html

E12. 'Translate at ExPASY.' http://expasy.hcuge.ch/www/dna.html

E13. 'ORNL Grail Form.' http://avalon.epm.ornl.gov/Grail-bin/EmptyGrailForm

E14. 'ExPASy Molecular Biology Server.' http://expasy.hcuge.ch/

E15. 'Harvard Biological Laboratory.' http://golgi.harvard.edu/

E16. 'Institut Pasteur.' http://www.pasteur.fr/

E17. 'What OLE Is Really About.' http://www.microsoft.com/oledev/olecom/aboutole.htm

E18. 'Distributed Object Computing with CORBA.' http://www.cs.wustl.edu/~schmidt/corba.html

E19. 'Information Resources for CORBA and the OMG.' http://www.acl.lanl.gov/CORBA/

E20. 'Sun's HotJava.' http://java.sun.com/

E21. 'JAVA on track.' http://www.javasoft.com/

E22. 'Javascript and Java.' http://home.netscape.com/eng/mozilla/Gold/handbook/javascript/introd.html

E23. 'Java/CORBA Working Group.' http://info.gdb.org/~letovsky/jcwg.html

E24. 'bioViews Consortium.' http://fruitfly.berkeley.edu/bioviews/

E25. 'Java-based Molecular Biology Work Bench.' http://www.embl-heidelberg.de/~toldo/JaMBW.html

E26. 'JADE home page.' http://alpha.crbm.cnrs-mop.fr/jade/jade.html

E27. 'PDB Viewer.' http://www.embl-heidelberg.de/~walther/JAVA/test2.html

E28. 'Human Physical Map.' http://genome.wustl.edu/cgm/cgm.html

E29. 'CINEMA.' http://www.biochem.ucl.ac.uk/bsm/dbbrowser/CINEMA/