A Survey of the Availability of Primary Bioinformatics Web Resources

Trias Thireou[1,2], George Spyrou[1], and Vassilis Atlamazoglou[1]*

[1] *Biomedical Informatics Unit, Foundation for Biomedical Research of the Academy of Athens, 11527 Athens, Greece;* [2] *Institute of Computer Science, Foundation for Research and Technology–Hellas, 71110 Heraklion, Crete, Greece.*

The explosive growth of the bioinformatics field has led to a large amount of data and software applications publicly available as web resources. However, the lack of persistence of web references is a barrier to a comprehensive shared access. We conducted a study of the current availability and other features of primary bioinformatics web resources (such as software tools and databases). The majority (95%) of the examined bioinformatics web resources were found running on UNIX/Linux operating systems, and the most widely used web server was found to be Apache (or Apache-related products). Of the overall 1,130 Uniform Resource Locators (URLs) examined, 91% were highly available (more than 90% of the time), while only 4% showed low accessibility (less than 50% of the time) during the survey. Furthermore, the most common URL failure modes are presented and analyzed.

Key words: bioinformatics resources, link validity, web reference persistence

Introduction

The World Wide Web has significantly improved the access to scientific information and resources, providing a way of making data and applications accessible and sharable, with the added convenience of information retrieval and extraction (*1*). The explosive growth of the bioinformatics field over the last years has led to a large amount of publicly available datasets and software applications. Web resources containing links for bioinformatics software tools and databases can be classified into five general groups: resource sites; news and discussions; publications; web lists and directories; and databases of biocomputing tools. Particularly, the members of the last group might serve both as long-term repositories of appropriate software tools as well as references for new software developers in order to avoid reinventing the wheel. Thus the Web has evolved from being of supplemental value to a primary resource for many scientists (*2–4*).

Although online resources have significantly aided scientific research, they also present new challenges to the traditional scientific process mainly due to the lack of persistence of web pages and sites. Studies concerning the increase and persistence of published Uniform Resource Locators (URLs) in research articles have been conducted in several scientific fields in order to identify to what extent URL decay is a problem in scientific literature (*5–7*). The aim of this study was to examine, quantify, and report the current availability and other features of open source or freely available applications under academic license, which are either provided as online web services or can be downloaded and executed locally.

Results and Discussion

Operating system

The selection of the type of operating system (OS) that will host a bioinformatics web resource is of significant importance. The distribution of the major OSs of the bioinformatics applications examined is demonstrated in Figure 1. According to this distribution, the vast majority of tools (95%) are running on UNIX/Linux systems, while a small percentage (4%) run on Win32 platforms, and an even smaller percentage (1%) is running on other platforms (Mac OS for example). This is because that UNIX/Linux systems have numerous attractive characteristics, such as complete development environment, networking facilities, high performance in terms of stability, speed, and

***Corresponding author.**
E-mail: vatlam@bioacademy.gr

availability of supporting software, as well as an active community of users/developers. Moreover, many computational biology and structural bioinformatics software applications have been developed for Silicon Graphics and Sun workstations.

The open source community has given a major boost in the use of Linux systems in many research fields including bioinformatics. Figure 2 describes the contribution of various Linux versions to the overall percentage of 48%. RedHat or RedHat-related Linux versions (CentOS, Fedora, and Mandrake) comprise 59% of the observed Linux versions, SUSE and Debian versions comprise 16% and 15% respectively, while the newly founded Darwin version since the year 2000 (Apple Computer, Inc., Cupertino, USA) has quickly gained respect.

URL availability

The URL availability was examined for 1,130 URLs of bioinformatics software tools and databases. Redirected links, which comprise 6% of the overall URLs checked, were also considered available. Figure 3 demonstrates that most of the sites (80% of the total URLs) were available 100% of the time during the survey. If a web page can be considered generally available when it is up at least 90% of the time, then 91% of the tested URLs were highly available, while only 4% of the sites had low accessibility (less than 50% of the time). We also examined the relationship between published and not published web resources in the scientific literature (Figure 4). Approximately 6% of the URLs referring to not published resources belong to the low availability range (less than 50% of the time), whereas the respective percentage of the URLs in the published class is 3%.

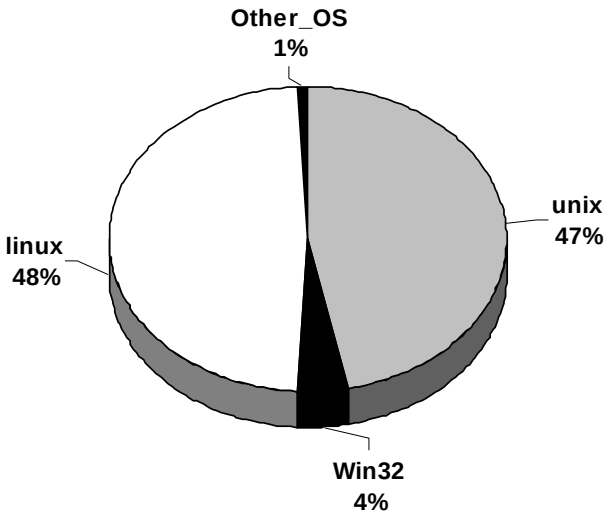

Fig. 1 Distribution of the major OSs for the examined bioinformatics tools and databases.

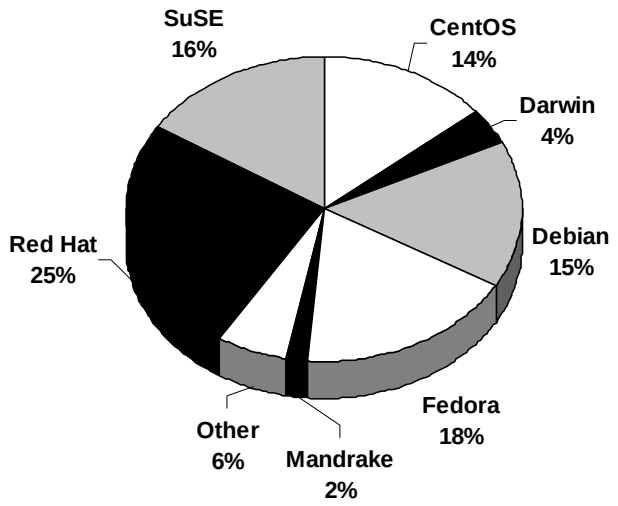

Fig. 2 Distribution of the major Linux versions used in the studied bioinformatics web resources.

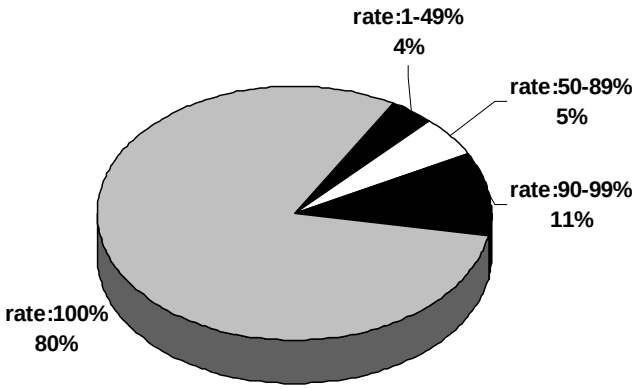

Fig. 3 Percentage of times that the tested URLs were available during the survey period.

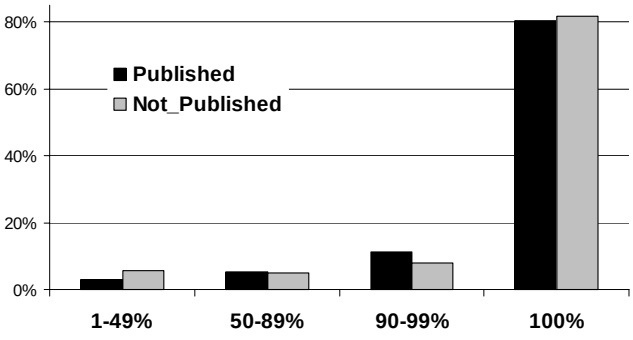

Fig. 4 Percentage of the published and not published bioinformatics web resources in various ranks of URL availability.

It is difficult to directly compare our results with those of previous studies, since the examined URLs correspond to primary resources (such as databases and applications) rather than secondary or supplemental web published documents. While the overall accessibility rate might be considered high compared with other studies (*3*, *6*), it should be reminded that the examined URL set is likely to exhibit slightly different characteristics. As it was previously mentioned, the vast majority of the tested URLs refer to bioinformatics applications published in scientific literature. Therefore, they are expected to have a relatively longer lifespan, since the peer-reviewed procedure takes into account the reliability/functionality and the suggested curation/maintenance protocols of the application. Furthermore, the majority of non-peer-reviewed published web resources have been collected from well organized and curated web lists and catalogs, containing software applications generally accepted and used by the scientific community that have passed the test of time. It is suggested that as a web page collection ages, it tends to be more stable (*5*).

Apache and Apache-related servers (such as Tomcat, Coyote, and Advanced Extranet Server) were found to support a large percentage (91%) of the total number of bioinformatics tools and databases. Apache is a full-featured open source web server with many powerful add-ons that runs on most commonly used platforms. It is interesting to note that in the case of less than 50% URL availability, 60% of the corresponding resources are supported by the Apache family, while for higher URL availability, the respective percentage is about 90% (Figure 5).

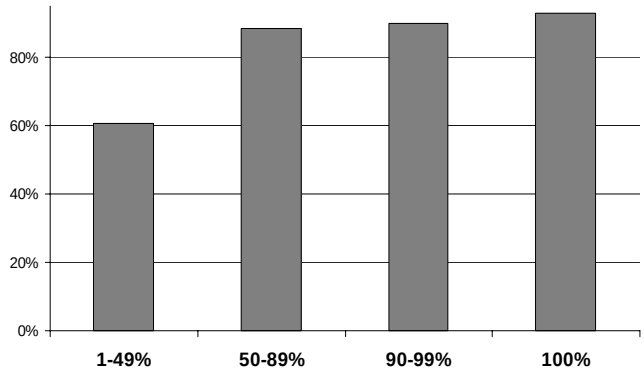

Fig. 5 Percentage of the bioinformatics tools and databases supported by Apache-related servers in various ranks of URL availability.

URL failure modes

For a web page to appear, a number of different technologies and protocols must work in concert, forming a complex chain of required actions. Any problem along this path would result in a failed URL request. Based on Figure 6, "cancelled_timeout", "connection_aborted", and "not_found" are the most prevalent failure modes for the examined dataset.

Figures 7 and 8 are an attempt to focus on server-related problems, which occur while resolving the host name and once the host is reached. By analyzing the data presented in Figure 7, it could be inferred that for the Windows OS, a failed URL reference is mainly related to "temporarily_overloaded" and "server_error" failure modes. On the other hand, "not_found" is the most common failure reason for UNIX, Linux, and the "other" OS type (with a relative frequency approximately three times as high as the one for UNIX/Linux). The relative contribution of "temporarily_overloaded" and "not_found" failure modes for the various Linux versions is shown in Figure 8.

Conclusion

The findings presented here have demonstrated the wide use of UNIX/Linux OSs and Apache-related technology. It is also suggested that primary resources exhibit higher availability rate compared with secondary or supplemental web published documents.

Materials and Methods

This study was conducted on a dataset of 1,130 URLs of bioinformatics software tools and databases from March 23, 2006 to July 21, 2006 on a weekly basis. The information was extracted from MetaBasis (http://metabasis.bioacademy.gr), a web-based relational database system for organizing and maintaining information relevant to bioinformatics tools and databases (*4*). These data mainly refer to open source or freely available applications under academic license. The vast majority (88%) of MetaBasis records have been published in scientific literature (Figure 9). From Figure 9 it can be seen that 70% of the studied entries have been published in the journals *Bioinformatics* (58%) and *Nucleic Acids Research* (12%). *Bioinformatics*, originally *Computer Applications in the Biosciences* in the 1980s, is the longest running

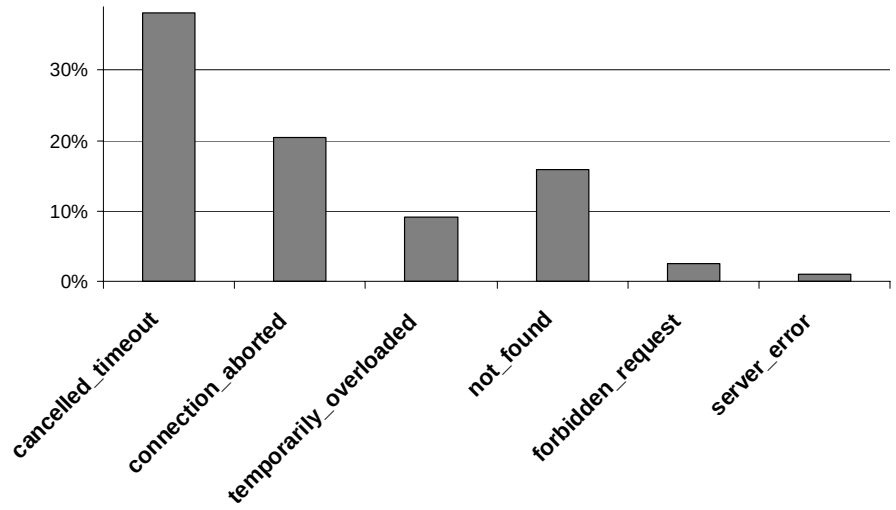

Fig. 6 Distribution of the broken URLs over the most common error types.

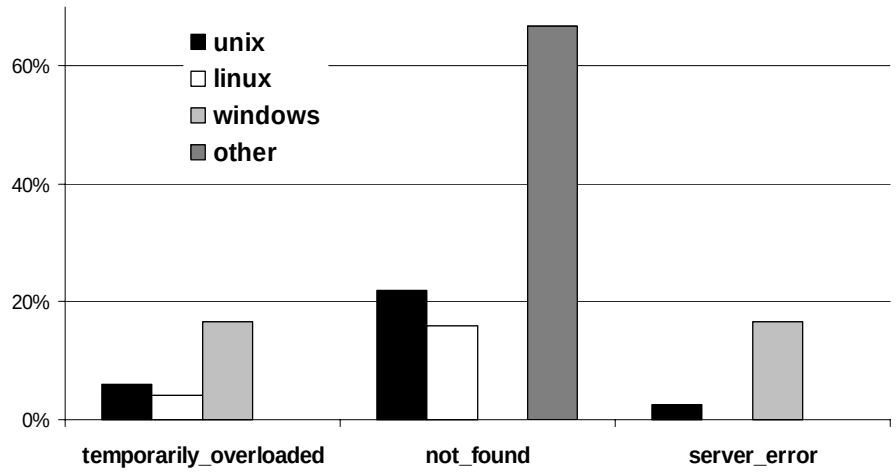

Fig. 7 Relative frequency of the most common server-related URL failure errors per OS.

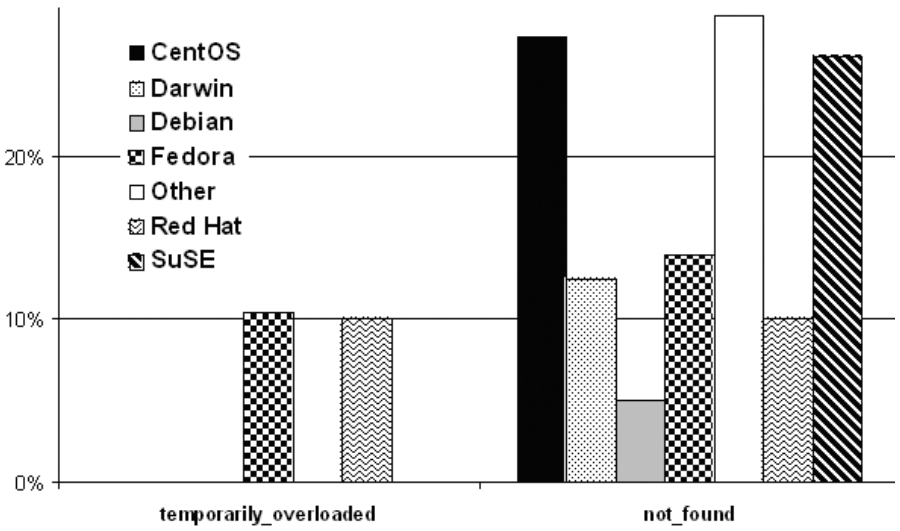

Fig. 8 Relative frequency of the most common server-related URL failure errors per Linux version.

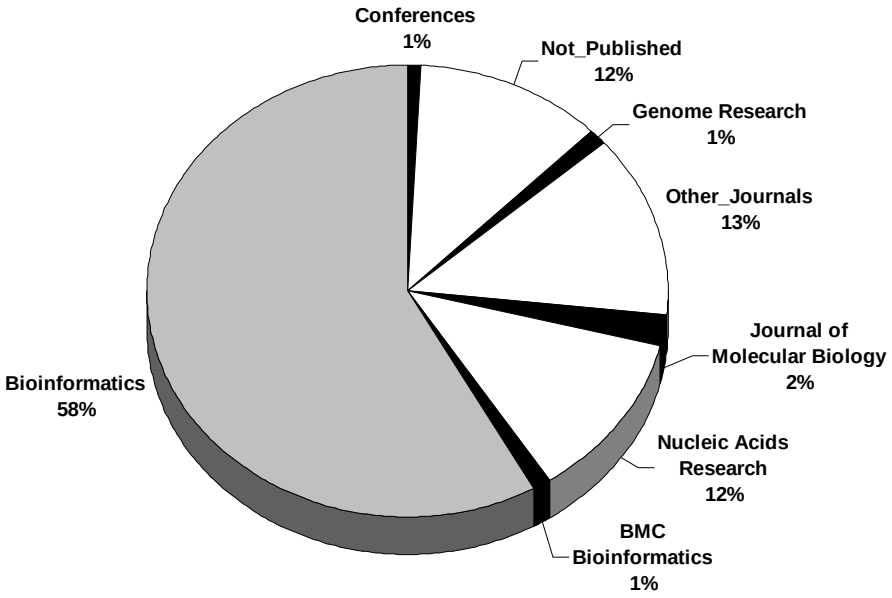

Fig. 9 Distribution of MetaBasis records across scientific literature. The vast majority was collected from scientific journals or conference proceedings.

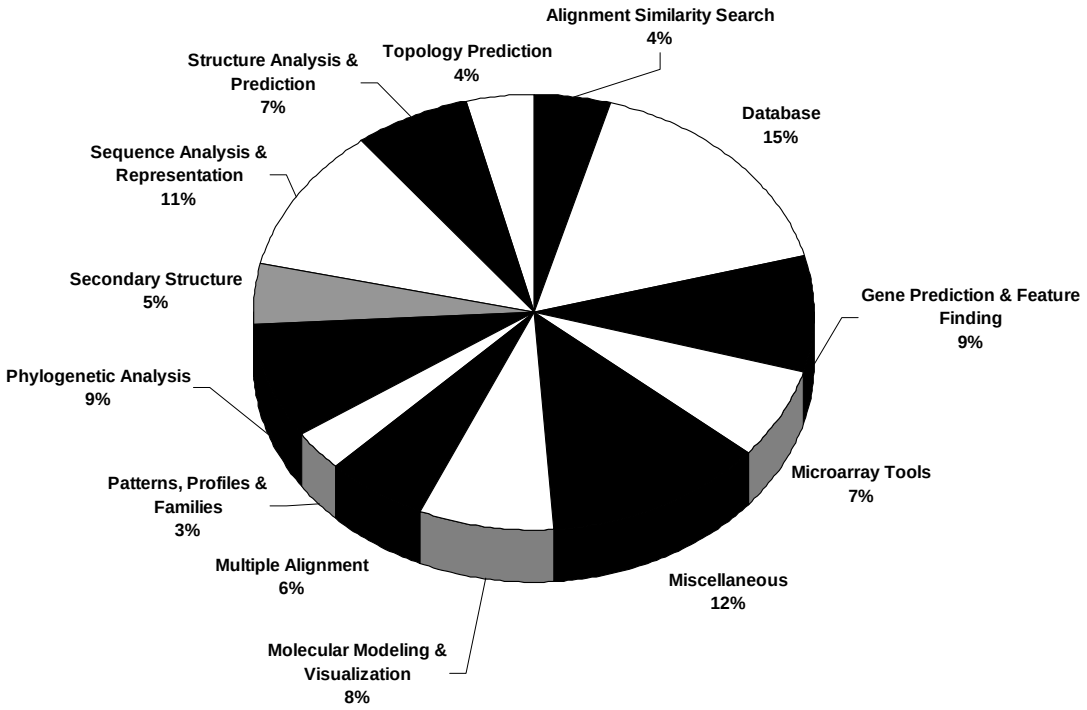

Fig. 10 Distribution of MetaBasis records across 13 predefined categories of bioinformatics tools and resources.

publication for original papers in this field, with a useful section of short application notes. Similarly in *Nucleic Acids Research*, the first issue of each year is devoted to biological databases, and an issue in July is devoted to papers describing web-based software resources of value to the biological community.

The examined records span a broad spectrum of available applications and resources in the field of bioinformatics, as it can be depicted in Figure 10. Furthermore, Figure 11 shows the distribution of the studied entries as a function of the time the resource was presented in scientific literature or the time it became available on the web, in case it was not peer-reviewed published.

Link (URL) validity check in a periodical basis, for all the deposited records, is part of the protocol followed for the curation of MetaBasis. Link reports were based on Xenu's Link Sleuth^{TM}and manual URL

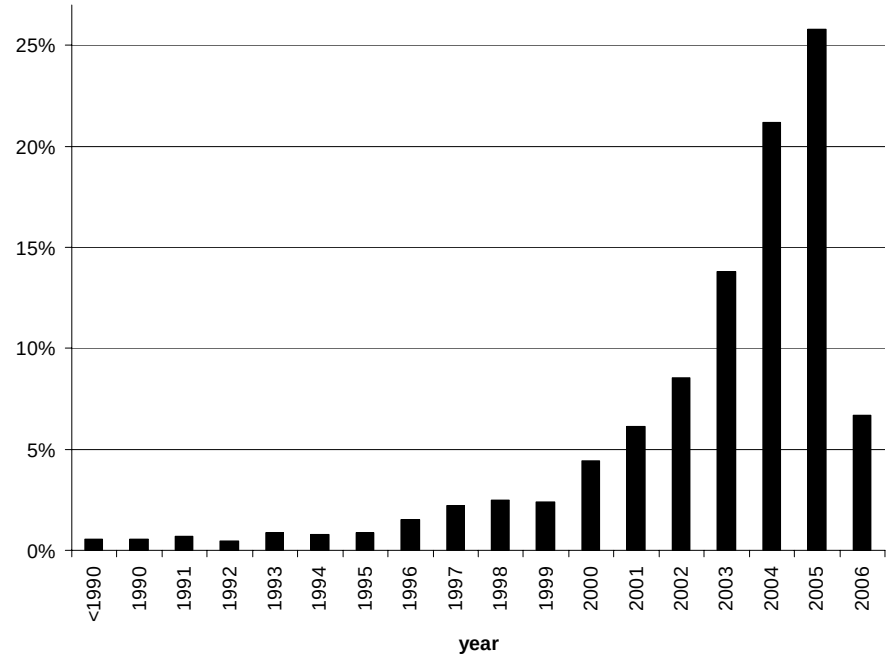

Fig. 11 The percentage of MetaBasis entries plotted as a function of publication time or application's web site creation.

Table 1 The main types of errors reported for broken links

Reported error message	HTTP response status code or WinInet error code	Textual description of the status
Forbidden request	403	The server understood the request, but is refusing to fulfill it.
Not found	404	The server has not found anything matching the requested URL. No indication is given of whether the condition is temporary or permanent.
Server error	500	The server encountered an unexpected condition which prevented it from fulfilling the request.
Temporarily overloaded	503	The server is currently unable to handle the request, due to a temporary overloading or maintenance of the server.
Cancelled/timeout	12017	The operation was cancelled or timed out, usually because the handle on which the request was operating was closed before the operation completed.
Connection aborted	12030	The connection with the server has been terminated.

validity check. Xenu's Link Sleuth checked web sites for broken links, and also detected and reported redirected URLs.

The list of the reported broken links was also checked manually in order to cope with the fact that some web sites might be programmed only for specific web user agents (such as Netscape and Internet Explorer) but refuse others. This procedure is also useful to deal with temporary network errors. The main types of errors reported for the broken links are tabulated in Table 1.

Additionally, a number of other features were also examined, such as the server type, the operating system used, and the date of publication or web site creation.

Authors' contributions

TT participated in the design of the study, the collection and processing of data, and the preparation of the manuscript. GS participated in the design of the study and the collection of data, and helped to draft the manuscript. VA conceived of the study, participated in its design and coordination as well as the collection of data, and drafted the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Lawrence, S. and Giles, C.L. 1999. Accessibility of information on the web. *Nature* 400: 107-109.
2. Gilbert, D. 2004. Bioinformatics software resources. *Brief. Bioinform.* 5: 300-304.
3. Wren, J.D. 2004. 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics* 20: 668-672.
4. Atlamazoglou, V., *et al.* 2006. MetaBasis: a web-based database containing metadata on software tools and databases in the field of bioinformatics. *Appl. Bioinformatics* 5: 187-192.
5. Koehler, W. 2002. Web page change and persistence—a four-year longitudinal study. *J. Am. Soc. Inf. Sci.* 53: 162-171.
6. Spinellis, D. 2003. The decay and failures of web references. *Commun. ACM* 46: 71-77.
7. Lawrence, S., *et al.* 2001. Persistence of web references in scientific research. *IEEE Comput.* 34: 26-31.