Characteristics and evolution of the ecosystem of software tools supporting research in molecular biology

Florencio Pazos and Monica Chagoyen

Corresponding author. Florencio Pazos, Computational Systems Biology Group, Systems Biology Program, National Centre for Biotechnology (CNB-CSIC), Madrid 28049, Spain. Tel.: +34-915854669; Fax: +34-915854506; E-mail: pazos@cnb.csic.es

Abstract

Daily work in molecular biology presently depends on a large number of computational tools. An in-depth, large-scale study of that ‘ecosystem’ of Web tools, its characteristics, interconnectivity, patterns of usage/citation, temporal evolution and rate of decay is crucial for understanding the forces that shape it and for informing initiatives aimed at its funding, long-term maintenance and improvement. In particular, the long-term maintenance of these tools is compromised because of their specific development model. Hundreds of published studies become irreproducible de facto, as the software tools used to conduct them become unavailable. In this study, we present a large-scale survey of >5400 publications describing Web servers within the two main bibliographic resources for disseminating new software developments in molecular biology. For all these servers, we studied their citation patterns, the subjects they address, their citation networks and the temporal evolution of these factors. We also analysed how these factors affect the availability of these servers (whether they are alive). Our results show that this ecosystem of tools is highly interconnected and adapts to the ‘trendy’ subjects in every moment. The servers present characteristic temporal patterns of citation/usage, and there is a worrying rate of server ‘death’, which is influenced by factors such as the server popularity and the institutions that hosts it. These results can inform initiatives aimed at the long-term maintenance of these resources.

Key words: computational method; Web server; server decay

Introduction

Molecular biology is increasingly becoming a data-intensive discipline. Although computers are essential nowadays in all scientific areas, their importance in molecular biology is greater because of the irreversible data-oriented trend in this field [1]. As a consequence, bioinformatics and related disciplines are becoming ever more important for dealing with the problems associated to the large amounts of data, i.e. how to handle, store and extract useful information from them [2].

Many bioinformatics methods and protocols, when sufficiently mature in terms of performance and usability, are implemented in computer tools that can be used by experimental molecular biologists. In most cases, a friendly Web interface is an important step in bringing these tools closer to users not expert in bioinformatics [3]. Web-accessible computational tools form part of today’s standard toolbox for all molecular biologists. In recent decades, there has been an explosion in the number of these Web-accessible tools as well as in the subjects they cover, and specific sections appeared in the scientific literature to publish and disseminate them [4].

While day-to-day work in molecular biology depends increasingly on these Web servers and programs, in many
cases, their long-term availability is compromised [5–8]. This is because the development model for scientific software is different from that of standard commercial packages. This software is usually implemented by PhD students or temporary workers, and its long-term maintenance is endangered by the mobility inherent to the scientific career [9]. Moreover, the current funding schemes in science are not well suited for long-term maintenance of such software, as they are objective-based and fixed-term. In most cases, these tools are developed in the context of a specific short-term project [10], and it can be difficult for a laboratory to maintain them in the long term. Even key bioinformatics resources such as KEGG [11] and Swissprot/UniProt [12] faced this funding problem, which threatened their availability. If such core servers, used daily by thousands of researchers, had these problems, it is easy to imagine the situation for smaller, more specific resources. The availability of these tools is nevertheless important, not only for future studies but for ensuring the reproducibility of past studies that used them. The loss of a Web server means that all protocols and studies that used it become irreproducible in practical terms, while reproducibility is a key factor in scientific advance. This problem has only recently been taken into consideration, although no clear solutions have been found [13, 14]. Formal initiatives were created to ensure long-term availability of core Web resources, mainly databases; this is one objective, for example, of the ELIXIR project [15]. But even if these emerging initiatives are extended to Web servers, for obvious reasons, they can only maintain a limited number of resources, and objective criteria are hence needed to decide which ones.

Previous works attempted to study and quantify this ‘server decay’ problem using different sets of servers, and related their fate with different characteristics [5–8]. The large number of Web servers available today allows this problem to be addressed quantitatively. In addition, it permits the study of other characteristics from a systemic point of view (e.g. subjects addressed by the servers, usage/citation profiles of the articles describing them), as well as relating both issues (i.e. get insight into the characteristics that determine the long-term fate of a server).

Here, we present a large-scale study of the characteristics of >5400 Web servers published over a period of >15 years in the ‘Application Notes’ section of the journal Bioinformatics (BiAN) and the annual ‘web server issue’ of Nucleic Acids Research (NARWS), the two main bibliographic resources for informing the community of new software developments in this area.

Results and discussion
Overall citation figures

A total of 5428 journal articles describing Web servers and software tools were analysed (1603 for NARWS and 3825 for BiAN; Figure 1A and B shows the distribution of the total number of citations depending on publication year. Older articles obviously tend to have more citations, as the corresponding tools have been available for longer. NARWS articles are more-cited overall (~20 citations on average compared with ~10 for BiAN), which might be explained in part by the distinct journal policies: Bioinformatics publishes many Application Notes each year, while Nucleic Acids Research is more ‘selective’ in its single yearly issue devoted to Web servers. Servers within the first quartile (Q1) of the distributions reach around 40 citations for NARWS and 30 for BiAN, and there is a large number of servers with hundreds of citations (outliers in the box plots in Figure 1A and B).

Some of these servers/tools were published but have never been cited. The proportion of servers with no citations (according to CrossRef -See Methods) published each year is shown in Figure 1C. The larger proportions in recent years are not meaningful, as these servers are still ‘young’. BiAN doubles NARWS in this proportion of never-cited servers (7 versus 3% on average). Both values are nonetheless low, meaning that the vast majority of servers are not forgotten but are used to obtain results or to develop related methods/software.

We evaluated the relationship between the number of citations of the servers and their hosting organization, as extracted from the last and penultimate parts of their URL domains. For BiAN, servers hosted in academic sites (e.g. .edu, .ac.uk) tend to have more citations than the average, while servers hosted in github have less. The detailed results are available as Supplementary File 4.

Figure 1. Citations for the BiAN and NARWS articles published each year (A and B). Box-plot representations of citation distributions for NARWS (A) and BiAN (B). (C) Fraction of articles published each year with 0 citations, in both journals.
Yearly citation profiles

In addition to these global figures for total citation numbers, the yearly citation patterns of individual servers are also informative. The relative citation profiles for all servers analysed in the two journals, grouped in nine clusters, are shown in Figure 2. The clusters group servers with similar citation profiles and represent different usage/citation behaviour over time. Most clusters are similar for the two journals. Some clusters represent servers of increasing popularity/usage. For example, Cluster #4 (146 servers published in BiAN and 61 in NARWS) consists of resources that have undergone an exponential increase in their usage/citation patterns. The number of citations increases slightly (or remains constant) the first 5 years and then boosts. Both of these clusters are enriched in similar MESH terms, ‘codon’ and ‘codon/genetics’ (in BiAN) and ‘codon’ (in NARWS). These and other keywords (Supplementary Figure S4) led us to interpret these clusters qualitatively as related to DNA, genetics and genomes. Cluster #8 (108 servers BiAN, 48 NARWS; see Figure 2 legend) also represents increasing citations but apparently reaching a plateau; the servers are much-cited initially, but this rate decays with time and appears to approach a steady state. The enriched keywords are ‘models, chemical’ and ‘amino acid substitution’ for BiAN, and none for NARWS. As a result, Cluster #8 is apparently protein-related: structure (secondary and 3D), sequence alignment and evolution.

We also found profiles with clear decreasing trends. Cluster #3 (86 servers BiAN, 45 NARWS) consists of servers with many citations in their first 1–2 years of life, followed by a sharp decrease, with a stable citation profile over the following 6 years. The enriched keywords, such as ‘protein/genetics’ (BiAN) and ‘gene expression regulation’ (NARWS), led us to think that these servers deal with expressed sequence tags, gene expression profiling and protein sequence alignments. Clusters #5 and #7 (76 and 120 servers, respectively, for BiAN; 67 for NARWS) also show decreasing citation patterns (that continue after 10 years), starting with a brief increase (popularity) at the outset. This popularity peak differentiates these clusters; it lies around Year 2–3 for #7 and NARWS, and Year 4 for #5. The pattern of enriched MESH terms produces no clear picture for these clusters, although they might be related to technical issues such as signal processing, signal integration and computer communications.

The remaining clusters represent citation profiles with ups and downs (valleys and peaks) in distinct time frames. For example, in Cluster #1 (54 servers BiAN, no equivalent in NARWS), servers receive many citations when published, then begin a decay that reaches minimum in the fifth year from which they gradually recover to finally reach their initial citation level in the 10th year. The only enriched MESH term for this cluster (‘proteins/analysis/chemistry/genetics’) is not informative. Cluster #2 (91 servers BiAN, 48 NARWS) shows the opposite pattern: citations begin to grow immediately after publication, reach a maximum at Year 5 and then decay. Their enriched MESH terms are related to ‘histocompatibility antigens’, ‘3D imaging’ and ‘protein conformation’ for BiAN (no clear terms for NARWS). Cluster #6 (80 servers BiAN, 16 NARWS) is enriched in ‘amino acid motifs’ and ‘gene expression profiling’ for BiAN and ‘protein structure’ for NARWS. Cluster #9 (64 servers BiAN) is enriched in keywords related to text mining and similar, such as ‘natural language processing’ and ‘controlled vocabularies’. Finally, an interesting cluster of NARWS servers (labelled #R in Figure 2), with no BiAN equivalent, is enriched in RNA-related terms, including ‘small interfering RNA’. The citation profiles for this cluster show a sharp boost just after publication, followed by a slight decay (Year 5) to a stable citation plateau.

The lists of enriched terms for all these clusters are available as Supplementary Figure S4.

Yearly preferred topics

To evaluate whether specific subjects/topics are addressed preferentially by computational tools published in a given year, we evaluated the enrichment of the MESH terms associated to articles published that year relative to all articles published by the same journal over the whole period studied (see ‘Methods’ section).

As expected, there were few enriched terms for NARWS. This is because the single annual issue of this journal devoted to Web servers is intended to cover all topics and is hence much less influenced by trendy subjects or hot topics. For BiAN, there is a much richer repertory of year-enriched keywords. ‘Word cloud’ representations of the enriched terms for each year are shown in Supplementary File 2. For each year, terms with an enrichment score (P-value of the hypergeometric test; see ‘Methods’ section) ≤1E-4 are shown, and their font size is proportional to the negative logarithm of that score. A global trend
is clear from protein-related topics in the early 2000s to

genomics-related at the end of the period analysed. Early on,

there was also enrichment in terms related to technical issues

such as 'databases', 'database managing systems', and 'infor-

mation storage and retrieval'. At the end of the period (2012–15),

there is clear enrichment in terms related to human studies.

These trends describe a scenario in which there was empha-

sis at the outset (~2000) on methodological developments, and

the preferential subject of study was proteins (e.g. amino acid

sequences, 3D structures, models). This changed gradually in

favour of nucleic acids (genomics, transcriptomics, GWAS, mas-

sive sequencing) and more recently human studies, as these
data became available.

Server availability

The proportion of servers published each year in both journals

that are apparently functional today (see 'Methods' section) is

illustrated in Figure 3A. Obviously, the older the server, the

greater the chances that it is dead. Of servers published 15 years

ago (~2002), only half remain alive. For servers published a

given year (Y), the figure also plots the fraction that will die in

each of the following years, assuming a constant decay of

D/(2016–Y) servers/year, where D is the fraction of servers pub-

lished that year (Y) that are dead today. On average, approxi-

mately 3–5% of the servers published in a given year die in each

of the forthcoming years, a similar figure for BiAN and NARWS.

This parameter has improved slightly in recent years, which

means that modern servers are somewhat more stable. This

could be because of different factors, such as better hardware/

software, lower costs of maintenance or better trained develop-

ers (who are not professional programmers in many of these

academic servers). In absolute numbers, taking the two journals

together, around 15 servers of all those published in a given

year will die in subsequent years.

This temporal server decay is similar for both journals. This

indicates that publication of a server in a more 'selective' forum

such as the Nucleic Acids Research Web server issue does not

assure greater chances of survival or long-term maintenance.

We also evaluated whether the articles describing servers

that are alive or dead are enriched in certain MESH terms (see

'Methods' section). We found no significant enrichments
(P-value <1E-3), which means that, in general, a server's fate is

unrelated to its topic or the type of calculations it performs.

To evaluate the relationship between server alive/dead sta-

tus and its overall citation trend, we quantified the latter as the

slope of the regression line obtained from its yearly citation pro-

file (see 'Methods' section and Supplementary Figure S1). A pos-

itive value means that server citation increased with time. For

the servers published each year, we plotted the ratio of the

mean slope values for those alive today and those that are dead

(Figure 3B). This ratio is >1.0 for both journals for most years,

which means that servers that remain alive tend to have cita-

tion/usage profiles with larger increases over time than those

that are dead (Supplementary Figure S1). The pattern is similar

for the ratio between total numbers of citations (Figure 3C);

servers published a given year and alive today have approxi-

mately twice the number of citations as those published the

same year that are dead. To illustrate this, we show the citation

profiles for all servers published in the two journals in 2 repre-

sentative years, colouring differently those that remain alive

and those that are dead (Supplementary Figure S3). These

eamples show that 'living' servers tend to have more citations

and increase with time.

We evaluated the relationship between the alive/dead sta-
tuses of the servers and their origin/hosting organization, as

etracted from the last and penultimate parts of their URL

domains. For BiAN, servers hosted in centralized repositories

e.g. github, sourceforge, bioconductor) tend to be more avail-
able than the average. On the other hand, servers hosted in acade-

mic sites (.edu, .edu, .ac, .ac, <country>) tend to be more dead. The detailed results on domain analysis are avail-
able as Supplementary File 4.

Citation networks

All the computational tools we analysed form an ecosystem in

which some tools make use of others, extend them or have a
similar goal. This ecosystem should be reflected in the pattern of inter-article (inter-tool) citations, as well as in citations by external papers (assumed non-tools) to them. We studied this by constructing two networks, one with direct inter-server citations, and one linking servers if they are cited by the same external articles (see ‘Methods’ section and Supplementary Figure S2). Note that NARWS and BiAN articles are combined in these networks. The first network is available in Cytoscape format as Supplementary File 3; the second is too large to be visualized, although calculations can be performed on it, and it is available on request.

Approximately half of the servers studied (2844) are involved in the first network (inter-server citations), meaning that they cite other servers or are cited by them, and the other half are isolated tools, never cited by others. The largest connected component of this network encompasses 80% of its nodes (2265). This network has 3880 edges (inter-server citations).

The distributions for the in- and out-degrees (incoming and outgoing citations) show the typical power-law shape (Figure 4); a small number of servers are highly cited, whereas most servers receive 0 or a small number of citations (note that servers with 0 incoming citations are included in the network if they cite others). For the out-degree, this means that a small number of servers cites many others (review articles and similar), and most servers cite a small number of others. The 92 servers highly cited by others (in-degree ≥7) are equally distributed between the two journals and most are alive today. There is no clear enrichment in MESH terms, either for the highly cited (in-degree) or for the highly citing servers (out-degree).

We found 19 overlapping clusters with 10 or more nodes (servers) in this network (see ‘Methods’ section). The largest cluster is composed of 26 servers (3 BioAN and 23 NARWS), with a dense inter-citation pattern (Figure 5). Five of the servers within this cluster are non-functional according to our automatic test (see ‘Methods’ section). The enriched MESH terms for this cluster indicate that it includes servers devoted to the analysis of gene expression profiling data. The second cluster is composed of 24 servers (18 BiAN and 6 NARWS), including 7 dead servers, and its enriched keywords point clearly to roles related to protein structure (Figure 5). Specific topics can also be assigned to the other clusters based on their enriched terms (not shown).

The network that links servers if they are cited by the same external articles (Supplementary Figure S2) contains 3589 nodes (servers; 66% of those studied) linked by 37 027 undirected edges, whose weights represent the number of common citers for a pair of nodes. We found 67 clusters with 10 or more nodes in this network; Supplementary Figure S5 shows the enriched MESH terms for the five largest clusters. Cluster #1 is formed by 533 servers dedicated to analysis of gene expression profiling data, especially in the context of signalling networks. Cluster #2 is related to protein structure, #3 and #4 to protein sequences (e.g. alignment, evolution) and #5 to human genetic variation analysis.

Conclusions

We carried out a large-scale analysis of >5400 published computation tools for molecular biology. This analysis, based on the largest number of servers surveyed to date, outlines a large ecosystem of inter-related software that evolves with time, both in size and in characteristics. This ecosystem is interconnected, as inferred from the two citation-based networks, which indicates that new methods tend to build on existing ones, improving or complementing them, and that developers are aware of existing tools. Despite this overall interconnectivity, these same networks also show evident clusters of related tools that deal with similar topics. In both networks, the two main clusters are related to proteins and DNA studies, a thematic separation we see in many other parts of this study.

Figure 4. Distribution of incoming and outgoing connections in the network of inter-server citations.

Figure 5. The two largest clusters in the network of inter-server citations. Servers published in BiAN in red, those published in NARWS, blue. Servers are labelled with the PubMed ID of their corresponding publication. Servers dead today are shown as smaller nodes. The enriched MESH terms for the servers (articles) of each cluster are shown. The entire network in Cytoscape format is available as Supplementary File 3.
Our results show that the majority of tools are cited (used by molecular biologists to analyse data or built on by bioinformaticians). NARWS is more selective and publishes fewer papers that are more often cited. Clustering the servers published in both journals by their temporal citation profiles leads to groups related to certain topics/subjects. These, together with the topics enriched in the servers published each year (for BiAN), can be interpreted in terms of popularity/importance of certain subjects over time. In general, we found a temporal shift from protein-related themes to DNA/genomes, and later to human-related studies. This shows that, as anticipated, the tools developed adapt to the specific needs of molecular biology at a given time, and to the kind of data generated. To interpret these results, we must also consider that the MESH vocabulary has a certain historical bias and evolves with time, adapting to new techniques, concepts and types of data.

Owing to the particular software development/maintenance model of these tools, a large fraction of them become unavailable with time (~50% in 14 years, 70% in the entire 17-year period studied for BiAN). This figure is identical to that of Veretnik et al. [7] in a small-scale study of the availability of servers published in Nucleic Acids Research over a 4-year period: 14% died in 4 years (i.e. 50% in 14 years). Both numbers are more optimistic than those from another small-scale, in-depth study of availability of a number of servers published in Nucleic Acids Research [8], which estimated that 25% of servers become unavailable within 3 years (i.e. 50% in 6 years). In this latter study [8], not only server availability was assessed but also details of its function (i.e. availability of testable sample data), which could explain its more pessimistic figures.

The hosting organization (and country of origin) of a server seems to be related to its number of citations and availability as well. The most interesting observation is that servers hosted in academic sites are less stable but receive more citations than the average, while central repositories such as github ensure a much better software availability, but at the expenses of being less cited.

We found no difference in the ‘decay rate’ of the tools depending on the journal in which they were published or their topic/theme. The first observation (also found by Wren et al. in a recent study [6]) is especially noteworthy if one considers the Nucleic Acids Research requirement that compromises the author to maintain the tool for 5 years after publication. Based on our figures, >20% of published servers do not fulfil this requirement. We observed that, in general, servers that are alive today have more citations and increase with time versus those that are now dead. But cause and consequence are difficult to assess: Does the server cease to be cited after dying, or do its developers no longer maintain an uncited/unused tool? Tracking server availability over time, to detect when it becomes unavailable, will certainly answer this question. We are currently pursuing this issue for a future in-depth study.

Death of a server could be because of many factors impossible to disentangle based on our current data; these include development of an improved tool or version, lack of resources for maintenance or ‘turn-off’ because of lack of interest/usage. The small-scale study by Schultheis et al. [8], which included direct questioning of the authors, points to the second factor (lack of resources for maintenance) as mainly responsible. If this is the case, open software and centralized repositories could help to alleviate the problem. Even considering the limitations of the automatic procedure we used to evaluate the ‘availability’ of this large number of tools, we think these figures are representative of the current scenario. In any case, our figures and those from other studies indicate that this problem of server death is worrisome. Our results comprise the most comprehensive characterization and quantification of this phenomenon to date. These numbers and trends should inform current initiatives devoted to the long-term maintenance of resources [15]. The problem is serious, as the death of a tool means that all studies that used it become in effect irreproducible. As pointed out by Bourne et al. [10], usage pattern data are crucial for defining the funding strategies for computational resources. There is an obvious need to define metrics based on these usage patterns, as well as on other data, to estimate the impact, costs and benefits of each resource [16]. At least those tools with a large number of citations or a citation profile that increases with time (promising tools) should be funded or maintained by global initiatives independent of the authors.

Materials and methods

We obtained the list of all tools and Web servers published in the journals Bioinformatics and Nucleic Acids Research over long periods of time, directly from the Web pages of these journals. We parsed the tables of contents available on the journals websites and followed the links to the pages with the detailed information on each server’s article. For Nucleic Acids Research we retrieved 1603 articles published in the annual Web server issues (NARWS) from 2003 (number 31, issue 13) to 2016 (number 44, issue W1). For Bioinformatics we retrieved 3825 articles published in the ‘Application note’ sections (BiAN) of the issues from 1996 (number 12, issue 1) to March 2016 (number 32, issue 12). For each of these 5428 articles, we retrieved the main data (title, authors, abstract, URL or the tool they describe, etc.) directly from the journal website using a Web crawler, as well as the list of articles that cite them from the CrossRef resource linked in the same pages. We found a small number of articles with identical URLs between NARWS and BiAN: 57 cases (3.6% of NARWS, 1.6% BiAN). For the 5428 articles and those citing them, we retrieved the PubMed entries and extract their MESH terms. These data were retrieved on 4 October 2016.

To evaluate the over-representation of a given MESH term in a subset of articles relative to a background set [e.g. those published in a given year relative to all published in the same journal over the whole period studied, or those with a given citation profile relative to all published], we used a hypergeometric test. For a given MESH term, we calculate its P-value as:

\[ p-value = \sum_{i=0}^{r} \binom{R}{i} \binom{R-i}{s-i} / \binom{R}{s}. \]

Where R is the size of the background, r is the number of articles in the background with that MESH term, S is the size of the subset evaluated and s is the number of articles in the subset with that term.

For servers with at least a 10-year history since publication, we clustered their citation profiles (number of citations per year) in that 10-year interval using the k-means approach (with k = 9). To account for citation tendencies rather than the absolute number of citations, these numbers are converted to z-scores before clustering. Absolute numbers of citations are also studied in another context (see below).

To evaluate whether a server/tool is still ‘alive’, we automatically queried the URL given in the article and recorded the HTTP status code returned. We performed two consecutive attempts with a read timeout of 120 s and a connection timeout...
of 60 s. We considered the server available if its returned code was ‘OK’ (code 200), and not-alive/not-functional/dead for the following codes: ‘timeout’, ‘file not found’, ‘password failed’, ‘unable to resolve address’, ‘unable to establish SSL connection’, ‘forbidden’ and ‘server unavailable’. Outcomes different from these are considered ‘unknown’, and those servers were not included in the analyses related to server availability. Whereas this is a crude, simple way of assessing whether a server is functional, it covers some typical scenarios that could make the original URL non-functional while the server is indeed alive: e.g. URL redirection: a server moved to another site will be regarded as ‘OK’ as long as the new URL is functional. Even if we might have included some false positives and negatives with this simplification, the approach allowed us to scan a large number of servers (>5000), which would have been unfeasible manually. The server statuses were retrieved on 27 October 2016.

To quantify the overall citation trend of a server, we calculated the slope of the regression line for its citation profile (Supplementary Figure S1). A positive value indicates that citations tend to increase with time, a negative value would indicate decrease and a value close to 0, a tendency to remain stable.

For assessing the relationship between the servers’ first-level domains (.com, .edu, .<country>, …) and their number of citations, we carried out a two-tailed Kolmogorov–Smirnov test to compare the mean number of citations for the servers within a given domain with the mean for all servers. For assessing whether the servers within a particular first-level domain tend to be enriched in available/alive servers, we carried out a hyper-geometric test as described previously. The same procedure was applied for non-available/dead servers.

Two citation networks were constructed in which the nodes were the articles that reported the servers (Supplementary Figure S2). In the first network (inter-server citation network), the directed/unweighted edges indicate whether a given server cites another (the corresponding articles). In the second (external citation network), the undirected/weighted edges indicate the number of citations by external articles (assumed ‘non-computational’) that a pair of servers has in common. Network visualization and calculations were performed with Cytoscape v3.2 (www.cytoscape.org). To locate clusters in these networks (groups of nodes highly interconnected and sparsely connected to the rest of the network), we used the Cytoscape plugins (apps) CytoCluster [17] for the first (directed/unweighted) and ModuLand [18] for the second (undirected/weighted) networks. In both cases, plugins were run with their default parameters.

Supplementary Data
Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements
The authors thank the members of the Computational Systems Biology Group (CNB-CSIC) for insightful discussions and feedback. The authors also thank C. Mark for editorial assistance.

Funding
This work was partially funded by the Spanish Ministry for Economy and Competitiveness (SAF2016-78041-C2-2-R) to F.P., who also acknowledges the Spanish Ministry for Education, Culture and Sports and the Fulbright Programme for the ‘Salvador de Madariaga’ sabbatical stay award (FRX15/00319), with which this work was initiated.

References

Key Points
• The software available to molecular biologists forms a large ecosystem of interrelated tools whose size and characteristics evolve with time.
• These tools adapt to the particular needs of molecular biology at a given time and to the kind of data generated, with a clear temporal shift from protein-related themes to DNA/genomes, and later to human-related studies.
• The rate at which published tools become unavailable is worrying (half have disappeared within 14 years of publication).
• The disappearance of a tool is related to its citations (highly cited servers are more likely to survive) and the repository/institution where it is hosted, but not to its subject or the impact of the journal in which it is published.