ANALYSIS OF THE USABILITY AND ARCHIVAL STABILITY OF OMICS COMPUTATIONAL TOOLS AND RESOURCES

Serghei Mangul
University of California, Los Angeles
Flood of omics software tools

S. Mangul, Interpreting and integrating big data in the life sciences, preprint -- https://doi.org/10.7287/peerj.preprints.27603v1
Many bioinformatics tools are hard to install

Emily Madden @_emadden · Oct 27
The tweet I come across as I’m taking a break from trying to properly install and run a bioninformatics tool. #perfecttiming

Ran Blekhman @blekhman
Say you pick 100 random bioinformatics software tools -- how many will you actually be able to access, install, and run?

Our new paper: biorxiv.org/content/early/... twitter.com/serghei_mangul...
Is user-friendly scientific software an oxymoron?
Challenges to effective software development and distribution in academia

- Software written by researchers tends to be written with the idea that all users will know as much about the code as its original authors.
- Incentives in academia heavily favor the publication of new software, not the maintenance of existing tools.
- There is lack of protocols to check insatiability of published software tools in academia.
Software development in academia versus industry

- Industry receive considerably more resources for developing user-friendly tools.
- Companies efficiently distribute industry-produced software using dedicated company units or contractors.
- There is little reward for continuous, long-term development and maintenance of tools in academia.
Poorly implemented tools will ultimately hinder progress in big data-driven fields
Software crisis is a reproducibility crisis

- Limited software usability and archival stability of computational tools leads to (computational) reproducibility crisis
- Computational reproducibility is the ability to replicate published findings by running the same computational tool on the data generated by the published study
Data: archival stability of omics computational tools and resources

- 51,236 biomedical papers from PubMed
- 99 randomly selected bioinformatics tools

<table>
<thead>
<tr>
<th>Journal name</th>
<th>Number of URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature Biotechnol</td>
<td>180</td>
</tr>
<tr>
<td>Genome Medicine</td>
<td>352</td>
</tr>
<tr>
<td>Nature Methods</td>
<td>403</td>
</tr>
<tr>
<td>Genome Biology</td>
<td>904</td>
</tr>
<tr>
<td>BMC Systems Biology</td>
<td>912</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>3131</td>
</tr>
<tr>
<td>PLoS Comp. Biology</td>
<td>3226</td>
</tr>
<tr>
<td>BMC Bioinformatics</td>
<td>6840</td>
</tr>
<tr>
<td>BMC Genomics</td>
<td>7651</td>
</tr>
<tr>
<td>Nucleic Acids Research</td>
<td>13103</td>
</tr>
</tbody>
</table>
Analysis of archival stability

- We downloaded 51,236 open access papers via PubMed from 10 systems and computational biology journals (Raw data in XML format)
- We developed an approach to extract 36,702 software links from the downloaded papers and verify the archival stability of links
- Timeout links were manually verified
Archival stability of abstract URLs

The figure shows the percentage of links that are broken, timed out, accessible, and redirected over the years 2005 to 2017. The left graph has a bar chart with years on the x-axis and the percentage of links in categories on the y-axis. The right graph is a line chart with time in years on the x-axis and the number of links on the y-axis.
Effect of social media

![Bar chart showing the effect of social media on altmetric score and readership. The chart indicates that broken and accessible conditions lead to higher altmetric scores compared to the timeout condition, with significant differences marked by asterisks.](image-url)
GitHub is the most popular platform to host scientific code
Tools hosted on GitHub have <3% of unreachable links
Analysis of the software usability

- We have randomly selected 99 tools across various domains of computational biology.
- Assign one undergraduate student to install 10 tools or more
- Record total time, number of commands to install the tools
- Total human time to install 99 tools — 72 hours

Protocol to check the insatiability of the tool
Many tools are hard or impossible to install

Installation Difficulty

- 51.0% Easy Install (15 minutes or less)
- 27.6% Not Installed
- 21.4% Complex Install
Passed: no manual intervention
Manual interventions are time consuming.
Easy-to-install* tools have fewer undocumented commands.

*<15min
Bioconda tools were always installable.
Installable tools have more citations
Conclusions

◦ Host software on archaically stable services e.g. GitHub
◦ Provide easy-to-use installation interface and get more citations
◦ Tools wrapped in package managers (e.g. Bioconda) are 100% installable!

Editorial (Genome Biology) is available at https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1649-8
Preprint is available at https://www.biorxiv.org/content/early/2018/10/25/452532
Future work

Serghei Mangul @sergei_mangul · Oct 30

1/2 CALL FOR PARTICIPANTS: Given the interest in our study about bioinformatics tools usability, we want to work on the follow up a paper about the package managers for bioinformatics tools. Please reply in this tweet if you are interested to participate! @thmosqueiro @blekhman
Acknowledgment

Thiago Mosqueiro, Dat Duong, Keith Mitchell, Varuni Sarwal, Brian Hill, Jaqueline Brito, Russell Littman, Benjamin Statz, Angela Lam, Gargi Dayama, Laura Grieneisen, Lana Martin, Jonathan Flint, Eleazar Eskin, Ran Blekhman
We have prepared Jupyter Notebooks that utilize the raw data described above to reproduce the results and figures presented in our paper

→ https://github.com/smangul1/good.software

smangul@ucla.edu
http://www.sergheimangul.com/

THANK YOU.