Sustained software development, not number of citations or journal choice, is indicative of accurate bioinformatic software

Paul P. Gardner[1,2*], James M. Paterson[3], Stephanie McGimpsey[4], Fatemeh Ashari-Ghomi[5], Sinan U. Umu[6], Aleksandra Pawlik[7], Alex Gavryushkin[8], Michael A Black[1]

Abstract

Computational biology provides widely used and powerful software tools for testing and making inferences about biological data. In the face of rapidly increasing volumes of data, heuristic methods that trade software speed for accuracy may be employed. We are have studied these trade-offs using the results of a large number of independent software benchmarks, and evaluated whether external factors are indicative of accurate software.

We have extracted accuracy and speed ranks from independent benchmarks of different bioinformatic software tools, and evaluated whether the speed, author reputation, journal impact, recency and developer efforts are indicative of accuracy.

We found that software speed, author reputation, journal impact, number of citations and age are all unreliable predictors of software accuracy. This is unfortunate because citations, author and journal reputation are frequently cited reasons for selecting software tools. However, GitHub-derived records and high version numbers show that the accurate bioinformatic software tools are generally the product of many improvements over time, often from multiple developers.

We also find that the field of bioinformatics has a large excess of slow and inaccurate software tools, and this is consistent across many sub-disciplines. Meanwhile, there are few tools that are middle-of-road in terms of accuracy and speed trade-offs. We hypothesise that a form of publication-bias influences the publication and development of bioinformatic software. In other words, software that is intermediate in terms of both speed and accuracy may be difficult to publish - possibly due to author, editor and reviewer practices. This leaves an unfortunate hole in the literature as the ideal tools may fall into this gap. For example, high accuracy tools are not always useful if years of CPU time are required, while high speed is not useful if the results are also inaccurate.

[1] *Department of Biochemistry, University of Otago, Dunedin, New Zealand.*
[2] *Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand.*
[3] *Department of Civil and Natural Resources Engineering, University of Canterbury, Christchurch, New Zealand.*
[4] *Parasites and Microbes, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, 8 Cambridgeshire, CB10 1RQ, UK.*
[5] *Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark.*
[6] *Department of Research, Cancer Registry of Norway, Oslo, Norway.*
[7] *New Zealand eScience Infrastructure, 49 Symonds St, Auckland, New Zealand.*
[8] *Department of Computer Science, University of Otago, Dunedin, New Zealand.*
*__Corresponding author__: paul.gardner@otago.ac.nz

Background

Computational biology software is widely used and has produced some of the most cited publications in the entire scientific corpus [**?**, **?**, **?**]. These highly-cited software tools include implementations of methods for sequence alignment and homology inference [**?**, **?**, **?**, **?**], phylogenetic analysis [**?**, **?**, **?**, **?**, **?**], biomolecular structure analysis [**?**, **?**, **?**, **?**, **?**], and visualization and data collection [**?**, **?**]. However, the popularity of a software tool does not necessarily mean that it is accurate or computationally efficient, instead usability, ease of installation, operating system support or other indirect factors may play a greater role in a software tool's popularity. Indeed, there have been several notable incidences where convenient, yet inaccurate software has caused considerable harm [**?**, **?**, **?**].

Progress in the biological sciences is increasingly limited by the ability to analyse large volumes of data, therefore the dependence of biologists on software is also increasing [**?**]. There is an increasing reliance on technological solutions for automating biological data generation (e.g. next-generation sequencing, mass-spectroscopy, cell-tracking and species tracking), therefore the biological sciences have become increasingly dependent upon software tools for processing large quantities of data [**?**]. As a consequence, the computational efficiency of data processing and analysis software is of great importance to decrease the energy, climate impact, and time costs of research [**?**]. Furthermore, as datasets become larger even small error rates can have major impacts on the number of false inferences