A Bioinformatician, Computer Scientist, and Geneticist lead bioinformatic tool development - which one is better?

Paul P. Gardner[1]*

Abstract

The development of accurate bioinformatic software tools is critical for the effective analysis of complex biological data. This study investigates the relationship between the academic department affiliations of authors and the accuracy of the bioinformatic tools they develop. By analyzing a corpus of previously benchmarked bioinformatic software tools, we mapped tools to academic fields of corresponding authors and evaluated tool accuracy by field. Our results indicate that ”Medical Informatics,” categorized under ”Technologies,” outperforms other fields in terms of software accuracy, with a mean proportion of wins in accuracy rankings higher than the null expectation. Conversely, tools developed by authors affiliated with ”Bioinformatics” and ”Engineering” fields produce lower accuracy tools. After correcting for multiple testing, no result is significant ($p > 0.05$). Our findings show a distinct lack of an association between academic field and bioinformatic software accuracy, highlighting that the development of interdisciplinary software applications can be hosted by any aligned department with sufficient resources.

[1] *Department of Biochemistry, University of Otago, Dunedin, New Zealand.*

***Corresponding author**: paul.gardner@otago.ac.nz

Background

Much is made of departmental divisions within academia; These can denote research and teaching expertise [1], influence hiring decisions, access to funding, publishing and the training of students recruited for research projects [2]. However, interdisciplinary subjects such as bioinformatics break down the traditional barriers between departments and subject areas [3, 4, 5].

Bioinformatics, is an interdisciplinary field that fuses biology, computer science, and mathematics, and now plays a pivotal role in modern biological research [3]. The development of bioinformatic tools and software is critical for interpreting complex biological questions, such as what evolutionary, structural and functional analyses of genomic, transcriptomic, and proteomic data can tell us. The field of “bioinformatics” can include many overlapping research fields that include computational biology, biomathematics, biostatistics, medical informatics and other similar areas.

Bioinformatics began gaining traction with the advent of high-throughput sequencing technologies, necessitating robust computational tools to handle the large amounts of data generated [3, 5]. Departments specializing in bioinformatics emerged from biology, computer science, and engineering faculties, each contributing to the field’s growth.

Bioinformatics inherently requires a multidisciplinary approach. Interdisciplinary research allows for the integration of methods and perspectives from different disciplines, leading to novel insights and solutions that a single discipline might not achieve independently [6]. This is especially relevant in bioinformatics, where expertise in biological sciences is crucial for understanding the data, while computational skills are needed to develop and implement algorithms, and analyze and interpret data.

The **biological and health sciences,** in particular genetics, biochemistry and molecular biology provide essential domain knowledge, ensuring that the software tools developed are biologically relevant and accurate. Biologists can identify the critical biological questions and ensure that the computational tools are designed to address these questions effectively. However, biology departments may lack the advanced computational expertise required for sophisticated software development. We have grouped the biological and health fields into “**domain experts**” in the following analysis.

The **mathematical, engineering and computational sciences**, or the “**development experts**” as we have dubbed them here, contribute significantly by providing expertise in algorithm development, data structures, and software engineering principles. These skills are essential for creating efficient, scalable, and robust bioinformatic tools. Nonetheless, the challenge for computer science, engineering and mathematics or statistics departments lies in acquiring a deep understanding of biological domains necessary to ensure the relevance and accuracy of their software tools.

Departmental differences might matter in the development of bioinformatic software tools as they can reflect the the varying expertise, resources, and perspectives that different academic fields bring to the table. The development experts may excel in algorithm efficiency, mathematical modelling, data handling and software engineering. However, the domain experts may bring a deep understanding of biological questions, data interpretation and limitations, and possibly better curation of control datasets. As a result, the success of a tool may depend more on the integration of diverse expertise rather than the specific departmental affiliation of its developers. This blending of skills could neutralize any potential disparities between departments, leading to comparable outcomes regardless of the department of origin.

As the field of bioinformatics evolves, and the dependence on research software tools increases due to growing data volumes, this study seeks to determine whether the academic department of a corresponding author is associated with the accuracy of the bioinformatic software tools they develop. Specifically, we investigate whether the presumed subject expertise, as indicated by the author's departmental affiliation, has a measurable impact on the accuracy of these tools, using a benchmarked corpus of bioinformatic software tools to assess this relationship.

Results

We are interested in the relationship between the accuracy of bioinformatic software tools and academic fields of study. Using a published corpus of accuracy rankings for bioinformatic software tools, we have mapped software tools to academic fields and evaluated the relationship between software accuracy and academic field.

Previously benchmarked and collated accuracy ranks were obtained from a published supplement [7]. We mapped the corresponding author addresses for published software tools to a standardised list of specific "fields of study" [8], we further grouped these into higher level general fields, and broader areas of expertise based on the address details of each corresponding author. The number of tools representing each general and specific field is illustrated in Figure 1. The majority of bioinformatic software is produced by corresponding authors who list at least one of Genetics, Bioinformatics or Computer Science or similar departments as their primary address (Figure 1A). For general fields, the Biological Sciences produce the bulk of software tools, followed by the Computer Sciences (Figure 1B).

The mean proportion of wins (e.g. if tool 'A' outranks tool 'B', this is a win for tool 'A') and the corresponding Z-scores give a way to rank fields of study on the relative accuracy of bioinformatic software benchmarked between 2005 and 2020 relative to the expected number of wins for random groupings (i.e. $wins = 0.5$). The greater the proportion of wins, and $(-1) * Z - score$, then the more times tools from a particular department outperformed competing tools in independent benchmarks.

The specific and general fields that outperformed other fields is "Medical Informatics" which is a branch of "Technologies", the same 16 software tools fall into both fields. Five of which are different parameter options for the MAFFT multiple sequence alignment tool [9]. These have a mean proportion of wins of 0.70 and 95% confidence interval of $0.53 - 0.85$, which excludes the null of 0.5. The corresponding z-score is -1.88 and $P = 0.29$ after correcting for multiple testing.

At the other end of the spectrum is "Bioinformatics", ironically authors that list "Bioinformatics" in their address field appear to write less accurate software for bioinformatic applications. The mean proportion of wins was 0.43, and 95% confidence interval of $0.33 - 0.53$. The corresponding z-score is 1.20 and $P = 0.46$ after correcting for multiple testing.

The general field "Engineering" also had a low rank. The mean proportion of wins was just 0.34, and 95% confidence interval of $0.16 - 0.60$. The corresponding z-score is 1.25. This general field is made up of several smaller specific fields that includes "Bioengineering and biomedical engineering", "Computer engineering" and "Electrical and electronics engineering" which individually did not have more than ten corresponding software tools, so were excluded from the more specific analysis.

The remaining general and specific academic fields have confidence intervals that include the null value of 0.5, and relatively modest z-scores that range from -0.49 to 0.96. The p-values for each were greater than 0.05.

For the highest level field classifications of software development expert, biological domain expert or interdisciplinary expert each had similar mean proportions of wins (0.51, 0.49 and 0.46 respectively). The interdisciplinary experts had a lower z-score of -0.87, which corresponds to $P = 0.46$ after correcting for multiple testing.

Conclusions and Limitations

We have tested the assumption that the academic department subject reflects the quality of research software. Our findings found no significant association after multiple-testing correction between the biological domain expert, the software developer expert or the interdisciplinary scientist with software tool rankings on accuracy measures. Leading us to conclude that assumptions regarding academic department and bioinformatic software quality are likely false. The general and specific research fields yielded similarly negative results (Figure 2A).

Our earlier paper found that long-term commitments to updating software was the leading factor associated with accurate software tools. This current study complements that finding, showing that not only do citation metrics (e.g. journal impact factors and author H-index), tool age and tool speed, but also academic fields of inquiry are **not** associated with prediction accuracy.

We focus on software tool accuracy here [10]. While speed, usability and some features of software tools are important, in our opinion the primary concern for bioinformatic software tools is the accuracy of the results they produce. As poor predictions may have long-term consequences for our general research field.

Medical Informatics, under the broader category of ”Technologies,” is identified as the top-performing group in developing accurate bioinformatic software tools. The tools include a

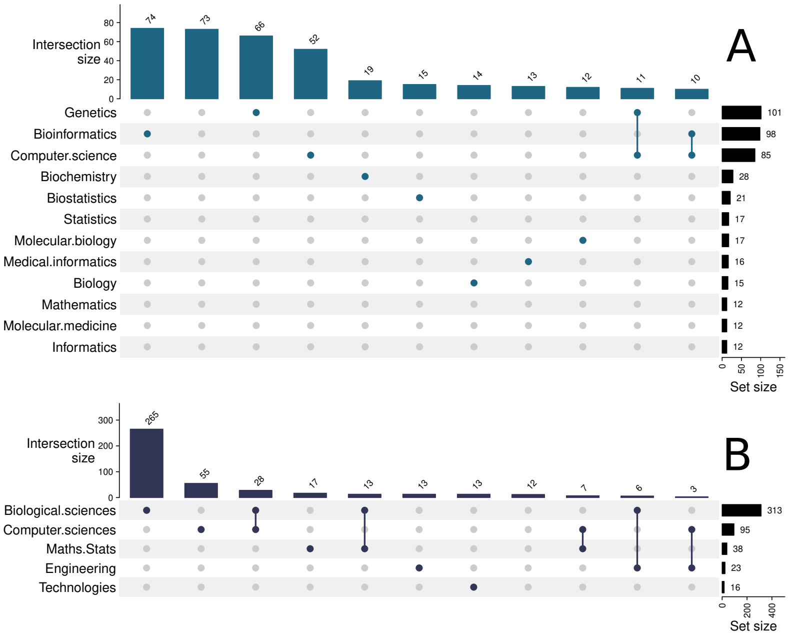

Figure 1. The number and intersection of general (**A**) or specific (**B**) fields that have contributed to bioinformatic software tools included in recent benchmark studies. The number and intersection of specific fields.

number of methods for structural variation detection, single-cell profiling, long-read assembly, multiple sequence alignment and are derived from several different research teams.

Bioinformatics and Engineering ranked lower in terms of software accuracy. Tools developed by authors who affiliated with ”Bioinformatics” typically had slightly lower accuracy than that of other fields. However, this was not a statistically significant finding. In addition (further bad news for this author), “Biochemistry” was similarly ranked, again this was not a statistically significant finding.

This leads us to conclude that an individual’s host department are not reflective of the quality of software that they are capable of producing ($p > 0.05$ in all instances). As a consequence, the academic department should not be used as a proxy for judging the potential of software development projects.

This study has several limitations. Some benchmarks rank several tool options, the effects of which may be modest to large, and potentially non-independent. The number of accuracy metrics used is also diverse, some of which may be flawed in some circumstances. For example, “accuracy” with large class-imbalance [11], or N50 which some commentators have criticised [12]. Some benchmarks are relatively small resulting in small changes in rank having a potentially large impact on the proportion of wins. Finally, the cohort of benchmarks has not been updated to include more recent results.

There is likely a disconnect between training and the host department for many bioinformatic tool developers. For example, this author studied mathematics as an undergraduate, conducted a PhD in bioinformatics, held postdoctoral fellowships in computer science and molecular biology departments, spent several years at a genomics research institute and is now employed in a Biochemistry Department. The “Biochemistry” departmental label is not an accurate reflection of my training or recent publications. I certainly can’t recall any of the key steps of glycolysis, the citric acid cycle or the structure of any amino acid or protein.

The last, corresponding author is generally the principal investigator leading a project. They may have limited involvement in the actual development of any software tool while they (supposedly) provide resources for the project and overall direction. However, there is a significant overlap between the department of a first author, who would generally be the primary developer of a tool, and the last author (not tested here). Therefore we expect these results to be broadly similar if first-author departments were used instead.

Methods

Pre-registration: This study’s desired sample size, included variables, hypotheses, and planned analyses were preregistered on the Open Science Framework prior to any unpublished data being collected [7].

Benchmarking data: software ranks from previously gathered benchmarks are publically available [13], these include data from 68 publications that rank the accuracy of different sets of 498 distinct software tools.

Mapping tools to academic field: For each software tool, when available, the corresponding publication(s) were identified and the addresses for the primary corresponding author

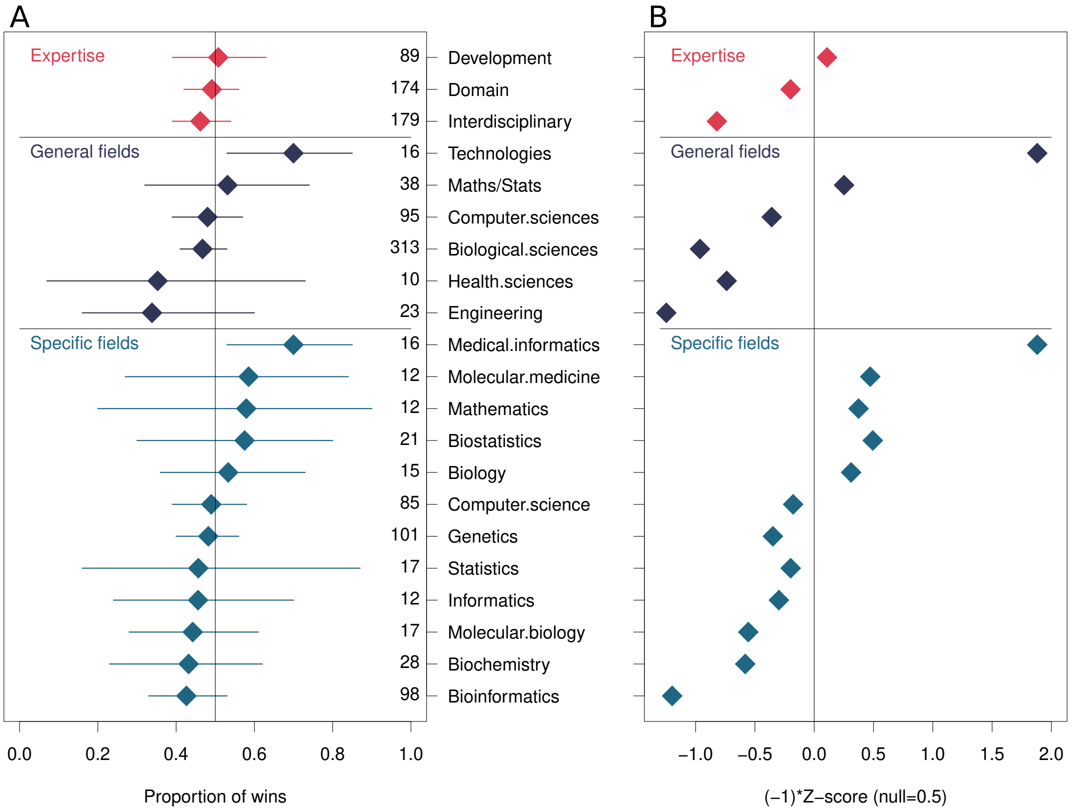

Figure 2. (**A**) A forest plot, illustrating the mean and 95% confidence intervals of the proportion of times software tools published by a given field "win" in pairwise comparisons. Confidence intervals and the mean was determined using a bootstrapping procedure. Within each field the entries have been sorted by the mean number of wins. The sample size for each field is indicated by the column of numbers on the right of the figure. (**B**) A Z-score was computed for each distribution of bootstrap samples for each field. The expected proportion of wins for randomly selected groups of tools was used as "*x*" (i.e. null=0.5).

were extracted manually. In cases where an author lists multiple addresses, just the first two are used. When multiple corresponding authors were listed, the last corresponding author was selected.

The author department names were mapped to the closest associated "fields of study" listed by the National Science Foundation [8]. We analysed fields at three different hierarchical levels, firstly the specific fields (e.g. "genetics", "computer science", "bioinformatics" etc). These are then mapped to more general fields within the "fields of study" classification (e.g. "biological sciences", "computer sciences" etc). We have further mapped these to three expertise types, the **development experts**, **domain experts** and **interdisciplinary experts**. The development experts are from fields such as computer science, mathematics and engineering, they are expected to bring development-pertinent expertise in software engineering and mathematical modeling of biological problems. The domain experts are drawn from the biological and health sciences and are expected to have detailed knowledge of the subject area and be invested in producing high-performing software for their research. The interdisciplinary experts are drawn from interdisciplinary subjects such as bioinformatics, biostatistics and biomathematics. Additionally the researchers who list both development and domain expertise in their addresses (e.g. "Computer Science" and "Genetics"). We have treated some fields as essentially synonymous, for example departments of "Computational Biology" was mapped to "Bioinformatics", and "Genomics" is mapped to "Genetics".

We restricted all subsequent analyses to fields that contain at least 10 software tools in our benchmark corpus. This mitigated against potential issues due to small sample sizes.

Statistical analysis: The accuracy data is derived from benchmarks using a diverse number of metrics that include sensitivity, specificity, PPV, FDR, error rates, AUROC, MCC and others [10]. The number of tools ranked in any benchmark ranged from 3 to 50. In order to obtain a representative measure of accuracy for a field that transcends diverse accuracy measures and numbers of ranked tools we elected to use ranks and a bootstrapping strategy. We randomly sample with replacement sets of 200 tools (from the total of 498), a benchmark that includes each tool is selected at random and the number of times the tool "won" against another tool is recorded, as is the total number of pairwise comparisons that were made. These counts of wins and total comparisons are transferred to the corresponding specific, general and expertise areas. This is repeated 1,000 times, and allows a way to estimate the mean proportions of wins allocated to each field, and a 95% confidence interval for this value (Figure 2A). We also compute a Z-score for each field using the below equation. This measures the number of standard deviations the mean number of wins is from the expected null value of 0.5 for randomly grouped tools (Figure 2B).

$z = \frac{x-\mu}{\sigma}$

Where μ is the mean, σ is the standard deviation, x is the raw value. In this case we set $x = 0.5$ as this is the null expectation for the proportion of wins for randomly grouped sets of tools. For the purposes of illustration we plot $(-1) * z$ so that the direction is the same as for the "proportion of wins" forest plot (Figure 1).

P-values are computed from the absolute value of the z-scores to evaluate if any field is significantly distinguished from the null i.e. $P[X > x]$. The P-values are corrected for multiple testing by controlling the false discovery rate method [14].

Data and analysis scripts availability

The data, scripts, figures and manuscript draft files are availble at the GitHub repository:
https://github.com/ppgardne/departments-software-accuracy

Funding: This research was supported by the MBIE data science platform "Beyond Prediction: explanatory and transparent data science", Genomics Aotearoa, and Marsden Grants 19-UOO-040 and 20-UOA-282.

Authors' contributions: this work was conceived by PPG, designed by PPG, the analysis was carried out by PPG, interpretation of the results was by PPG, the manuscript was drafted by PPG. All authors revised and edited the final manuscript. All authors approve the submitted version.

References

[1] Ben R. Martin. What's happening to our universities? *Prometheus*, 34:7–24, 2016.

[2] Paul Bourke and Linda Butler. Institutions and the map of science: matching university departments and fields of research. *Research Policy*, 26(6):711–718, 1998.

[3] C A Ouzounis and A Valencia. Early bioinformatics: the birth of a discipline–a personal view. *Bioinformatics*, 19(17):2176–90, Nov 2003.

[4] S R Eddy. "antedisciplinary" science. *PLoS Comput Biol*, 1(1):e6, Jun 2005.

[5] Paulien Hogeweg. The roots of bioinformatics in theoretical biology. *PLoS computational biology*, 7(3):e1002021, 2011.

[6] Fulvio Mazzocchi. Scientific research across and beyond disciplines: Challenges and opportunities of interdisciplinarity. *EMBO reports*, 20(6):e47682, 2019.

[7] P P Gardner. Pre-registration: Which department makes the best software?, 2024. `https://doi.org/10.17605/OSF.IO/92PTZ` [10 July 2024].

[8] IPEDS Completions Survey; National Center for Science Department of Education, National Center for Education Statistics and Survey of Earned Doctorates. Engineering Statistics. Classification of fields of study, 2014. `https://ncsesdata.nsf.gov/sere/2018/html/sere18-dt-taba001.html` [Accessed: July 2024].

[9] Kazutaka Katoh and Hiroyuki Toh. Recent developments in the mafft multiple sequence alignment program. *Briefings in bioinformatics*, 9(4):286–298, 2008.

[10] Lukas M Weber, Wouter Saelens, Robrecht Cannoodt, Charlotte Soneson, Alexander Hapfelmeier, Paul P Gardner, Anne-Laure Boulesteix, Yvan Saeys, and Mark D Robinson. Essential guidelines for computational method benchmarking. *Genome biology*, 20:1–12, 2019.

[11] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de Las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, 2019.

[12] Luyu Xie and Limsoon Wong. Pdr: a new genome assembly evaluation metric based on genetics concerns. *Bioinformatics*, 37(3):289–295, 2021.

[13] P P Gardner, J M Paterson, S McGimpsey, F Ashari-Ghomi, S U Umu, A Pawlik, A Gavryushkin, and M A Black. Sustained software development, not number of citations or journal choice, is indicative of accurate bioinformatic software. *Genome Biol*, 23(1):56, Feb 2022.

[14] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.