A metabenchmark of bioinformatics software

- Does bioinformatic software trade speed for accuracy?
- Metabenchmarking bioinformatics
- Is slow software more accurate?
- Benchmarking bioinformatics software
- Most bioinformatics software is slow and inaccurate

Paul Gardner[1,2,3,*], Fatemeh Ashari Ghomi[1,2], Sinan Uğur Umu[1,2], Stephanie McGimpsey[4]

[1]School of Biological Sciences, [2]Biomolecular Interaction Centre, [3]Bio-Protection Research Centre, University of Canterbury, Christchurch, New Zealand

[4]Dance Academy, Ireland

*paul.gardner@canterbury.ac.nz

Abstract

Introduction

Bioinformatics software is very widely used and has produced some of the most cited publications in the world (Van Noorden et al. 2014). In this work we use a liberal definition of the term “bioinformatics”, namely the application of mathematical and computational tools to study the structure and function of biological data. Bioinformatics software include methods for sequence alignment and homology inference (Thompson et al. 1994)(Thompson et al. 1997)(Altschul et al. 1990)(Altschul et al. 1997), phylogenetic analysis (Felsenstein 1985)(Saitou and Nei 1987)(Posada and Crandall 1998)(Ronquist and Huelsenbeck 2003)(Tamura et al. 2007), statistical analysis of survival patterns in biomedicine (Kaplan and Meier 1958; Cox 1972), biomolecular structure analysis (Sheldrick 1990)(Sheldrick 2008)(Otwinowski and Minor 1997)(Laskowski et al. 1993)(Jones et al. 1991), biomolecular visualization (Kraulis 1991) and storage (Berman et al. 2000). Yet the popularity of a computational tool or software suite does not necessarily imply that it is the most accurate or computationally efficient. The most appropriate method for determining this, when a choice is available, is to benchmark methods using both a mix of positive and negative controls. The tools that produce the best tradeoff in terms of false predictions, true predictions and speed may be well suited for the user. Sometimes these benchmarks are published and serve a useful role in reducing the “over-optimistic reporting” of bioinformatics accuracy (Boulesteix 2010)(Jelizarow et al. 2010).

Bioinformatics software is frequently said to trade accuracy for speed. For example, the classic homology search problem based upon sequence similarity has a mathematically optimal

solution in the Smith-Waterman algorithm, a dynamic programming method that finds optimal local alignments between two sequences. However, this approach is usually considered to be too slow for practical screening of large biological sequence databases, therefore heuristic methods such as BLAST are frequently used that trade the mathematical guarantee of an optimal solution in exchange for speed (Altschul et al. 1990)(Altschul et al. 1997).

Consequently, researchers that need to select the best available tool for a bioinformatic task, may select the slowest method. Since, in theory, this should be accurate. Or, they may select methods that are the newest and the current state-of-the-art, or the most popular or published in the highest profile journals.

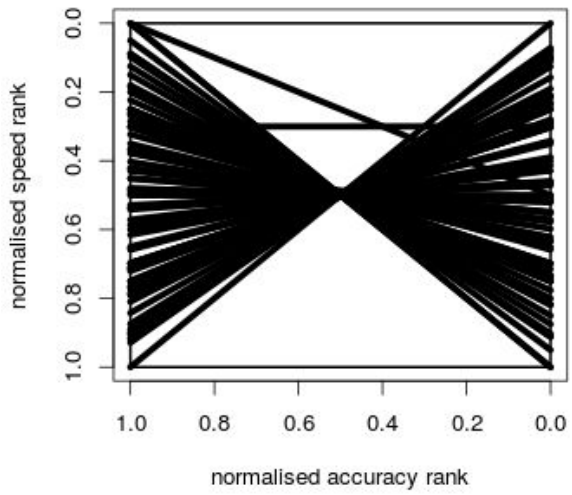
In this study, we investigate a) whether bioinformatic tools trade speed for accuracy and, b) whether the slowest, newest, highest profile or the most popular method is most likely to be the best.

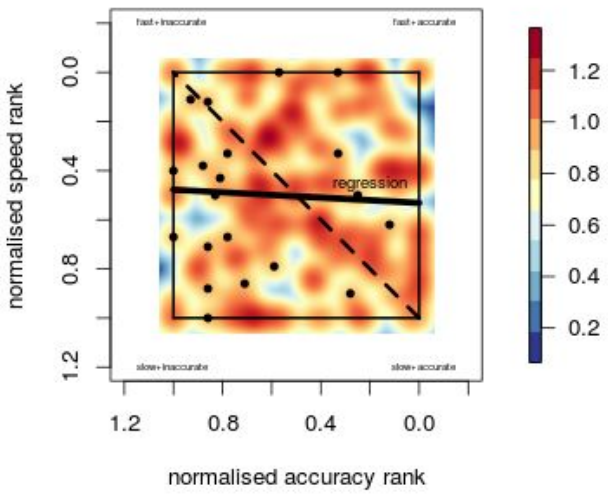
Points to make:

- a Cochrane Review for bioinformatics
- Frequent use of inefficient tools ruin the environment -- CPU cycles consume a lot of energy
-

Results

https://docs.google.com/spreadsheets/d/14xIY2PHNvxmV9MQLpbzSfFkuy1RlzDHbBOCZLJKcGu8/edit?usp=sharing





Discussion

- The most accurate methods are not always the slowest. They are rarely the fastest, however many are quick.
- Slow and inaccurate software is common, very fast and very inaccurate software is rare. May be a result of iterative software testing pushing the faster methods to be more accurate zone.
- Relationship with the impact factor of the publishing journal and/or the age of the method and/or the number of citations?
    - is the most cited method “the best”?
    - is the newest method “the best”?
    - is the method published in the highest impact journal “the best”?
        - Label methods as “high impact” that were published in high IF journals (IF>10) or receive > 50 citations per year. Are these above or below the line?
- A Cochrane Review for bioinformatic tools would be great!
- How many methods have acc+speed>1 → ie. slow and inaccurate?
- How many methods have acc+speed<1 → ie. fast and accurate?
- Fatemeh: contradicts some findings of game theory: e.g. resource abundance...

Conclusions

- Software development is usually an iterative and nonlinear process. Software is tweaked, tested and then refined. Faster methods can undergo more development cycles than slower software.
- When initiating a software development project, developers may find that using biologically reasonable fast and/or heuristic approaches produces a better result than beginning with a slow, more mathematically complete approach.
- Add that slow and inaccurate methods serve a useful purpose, like all negative results, they imply that certain reasonable sounding approaches may not work. They should not be discriminated against, furthermore, with more development they may result in an accurate (and/or fast) method.

Methods

Criteria for inclusion: We are interested in using bioinformatics benchmarks that satisfy Anne-Laure Boulesteix's (ALB) criteria for inclusion (Boulesteix et al. 2013). Specifically, (A) the main focus of the article is the comparison (not the introduction of a new method), (B) the authors should be reasonably neutral and (C) the test data and evaluation criteria should be sensible.

Literature mining: We identified an initial list of 10 benchmark articles that satisfy the ALB-criteria. These were identified based upon direct knowledge of published articles combined with several literature searches (e.g. '"benchmark" AND "cputime"' was to query both GoogleScholar and Pubmed (Sayers et al. 2010)). We used these articles to seed a machine-learning approach for ranking further candidate articles.

Data extraction:

Analysis

1. Compute a rank correlation for each benchmark,
2. Use a mixed model on the ranks to determine if accuracy is negatively correlated with speed.

Method

1. Develop a series of search terms for scanning bibliographic databases (E.g. Scopus, WoS, GoogleScholar, ...) for relevant articles. Filter these using machine-learning methods (e.g. Abstrackr)? Rank articles based upon the likelihood that they contain the results that we are after.

2. Search terms: benchmark methods algorithms genome data datasets challenge comparison evaluation performance accuracy
3. See supplementary material for: (Boulesteix et al. 2013)
4. 10 benchmark papers, 46 “benchmarks”, 74 methods, 371 datapoints.

Acknowledgments

Shinichi Nakagawa, Suetonia Palmer, Sinan Umu, Fatemeh Ashari Ghomi