Bioinformatics, 32(17), 2016, 2686–2691
doi: 10.1093/bioinformatics/btw284
Advance Access Publication Date: 5 May 2016
Original Paper

OXFORD

Data and text mining

Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades

Jonathan D. Wren[1,2,*]

[1]Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104-5005, USA, [2]Department of Biochemistry and Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on January 27, 2016; revised on April 1, 2016; accepted on April 21, 2016

Abstract

Motivation: To analyze the relative proportion of bioinformatics papers and their non-bioinformatics counterparts in the top 20 most cited papers annually for the past two decades.

Results: When defining bioinformatics papers as encompassing both those that provide software for data analysis or methods underlying data analysis software, we find that over the past two decades, more than a third (34%) of the most cited papers in science were bioinformatics papers, which is approximately a 31-fold enrichment relative to the total number of bioinformatics papers published. More than half of the most cited papers during this span were bioinformatics papers. Yet, the average 5-year JIF of top 20 bioinformatics papers was 7.7, whereas the average JIF for top 20 non-bioinformatics papers was 25.8, significantly higher ($P < 4.5 \times 10^{-29}$). The 20-year trend in the average JIF between the two groups suggests the gap does not appear to be significantly narrowing. For a sampling of the journals producing top papers, bioinformatics journals tended to have higher Gini coefficients, suggesting that development of novel bioinformatics resources may be somewhat 'hit or miss'. That is, relative to other fields, bioinformatics produces some programs that are extremely widely adopted and cited, yet there are fewer of intermediate success.

Contact: jdwren@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Background

Computational methods have long been important for data analysis and interpretation in biomedicine, rising steadily in their use over time and prevalence in most disciplines (Perez-Iratxeta *et al.*, 2007). The birth of bioinformatics, as a field, is difficult to trace to a specific date, but rather it began to emerge as soon as early computers did in the 1950s (Ouzounis and Valencia, 2003) and grew proportionally as access to computational technology increased. In the early days, it did not make much sense to think of bioinformatics as its own field, as computers were just tools for the analysis, organization and sharing of data. Often, whomever was most comfortable in a lab with using computers became an early bioinformatician by default. But as macromolecular data (e.g. DNA, protein) grew in both amount and diversity, and patterns began to emerge in terms of evolution by recombination and mutation rates (Koch, 1971; Ohta and Kimura, 1971), as well as the effects that mutations might have on the 3D structures of proteins, it became increasingly apparent that a specialized combination of skills would be required to advance understanding. Such data was rapidly becoming impossible to analyze *without* computational methods. Although formal training was not necessary, one would need to understand both biology and computer science—neither one alone would be sufficient, which is still true today. For example, although sequence alignments could be (and were) done manually, algorithmic methods (Needleman and Wunsch, 1970) would rapidly become indispensable, particularly as more data became publicly available. As strange as it may sound to the modern bioinformaticist, new sequence data was published in print form as late as 1984 (Anderson *et al.*, 1984).

In the absence of specialists to turn to, geneticists and biologists would learn what computational methods they needed to continue to advance their understanding of molecular biology. As the types of algorithmic approaches used became more applicable to different fields, such as finding common substrings (Aho *et al.*, 1976) or phylogenetic distance calculation (Doolittle, 1981), which could be used for analysis of macromolecules or text, for example, there began an increased interest in understanding how computational methods could advance our understanding of biology and medicine in ways that had not yet been anticipated. For example, could protein-coding regions be predicted solely from genomic sequence data? Because research into such methods would often be somewhat of tangential interest to biomedical and computer science journals alike, specialized journals would be needed. In 1985, *Computer Applications in the Biosciences* (CABIOS), the predecessor to the journal *Bioinformatics*, was one of the first to focus on this new area (Beynon, 1985; Sander, 2001).

The advantage to publishing in a cross-disciplinary journal is that it enables both recognition and reporting of novel research that might otherwise be rejected from either parent field alone. The disadvantage is that such papers often receive less interest from either of their parent fields alone. The impact of some of bioinformatics 'superstar' papers (i.e. those that receive the highest number of total citations among all papers published within a period of at least a year) has been anecdotally discussed before, such as BLAST (Van Noorden *et al.*, 2014) and SHELX (Schwarzenbach *et al.*, 2010), but the goal of this report was to examine how often papers reporting the development of novel bioinformatics methods or software achieve a disproportionately high impact ('superstar' status) among all papers published in science during the same year, and whether or not the potential impact was evident at the time. The assumption made for the latter estimate is that authors generally try to publish in higher impact journals whenever possible, although other considerations such as journal suitability will also factor into their decisions. If superstar papers in a specific field tend to be disproportionately published in journals with lower journal impact factors (JIF), then it is reasonable to hypothesize that the editors, reviewers and/or possibly the authors did not foresee the impact the paper would have on science.

Another motivation for this study is that recent studies have highlighted problems bioinformaticians face in being formally recognized. For example, bioinformaticians often are hired or funded to play more of a collaborative role within the research of other scientists, which creates challenges in terms of career advancement because traditional metrics tend to reward the primary drivers of projects (Chang, 2015). Bioinformaticists are often middle authors on papers, and prior studies have shown that, as the number of authors on the byline grows, the perceived contributions of the middle authors drops rapidly (Wren *et al.*, 2007). Also, another study found that while two-thirds of the biology papers they analyzed mentioned the use of software, less than half actually cited it (Howison and Bullard, 2015). Similarly, another study found that very few of the users of ArrayExpress properly cited it (Rung and Brazma, 2013) suggesting that the impact of bioinformatics software may be underestimated by citation metrics.

Given that key authorship positions on papers are valued disproportionately and that JIF is commonly used as a metric to evaluate the relative importance or impact of one's work, whether the actual impact of primary bioinformatics research correlates with its perceived impact at the time of publication is important to know. Or in other words, are important bioinformatics solutions generally recognized to the same degree that important biological findings are when they are first sent for publication?

To answer this question, citations are used as a proxy for importance, largely because they are quantitative, non-controversial and readily obtainable. Since the degree of importance can only be objectively estimated in retrospect, historical data is used. And, because the concept of citation-based 'importance' is relative (i.e. the number of citations to a paper is only interpretable within the context of a reference group), all papers published during a given year were used as the reference point because they have had approximately the same amount of time to accrue citations.

Finally, for a time, the number of popular references to 'bioinformatics' was rapidly on the rise, as the potential of the field was being recognized, but such references have recently been on the decline (Ouzounis, 2012). One might suggest perhaps bioinformatics did not live up to the initial hype, but there is usually 'term fatigue' associated with new words and also as bioinformatics approaches become ubiquitous, it may be merging back into the parent fields from which it was born. Nevertheless, this trend change suggests it is an appropriate time to examine the question: How much of an impact has bioinformatics had upon science?

2 Methods

2.1 Journal impact factors and citations

Statistics on citations were downloaded from the Institute for Scientific Information (ISI)'s Web Of Science (http://apps.webofknowledge.com/) using their web interface on March 21, 2016. Review papers were excluded and results were sorted by the number of citations found in the Web of Science Core Collection, The default number of citations provided by ISI includes citations from datasets (from their Data Citation Index), which sometimes is highly disproportionate to their influence as measured by literature-based citations. For example, PMID 23470992 only has 24 citations from papers, but 8697 from deposited datasets, making it the most cited paper of 2013 by that metric.

To calculate journal impact factor (JIF), 2013 5-year impact factors were used (the average number of times articles from a journal published in the preceding five years were cited in 2013). The 5-year JIF was used because it should fluctuate less than the regular 2-year JIF and be more reflective of long-term trends. Although JIFs change over time, the general trend is towards a gradual increase in all JIFs (Althouse *et al.*, 2009), so to control for this, the JIF was pegged to a fixed reference point (2013), which is similar to the approach used in economics to control for the changing value of fiat currencies. All but 5 journals were mapped to 2013 5-year impact factors. Journals that underwent name changes (e.g. *CABIOS* became *Bioinformatics*) were mapped to the current name. For journals that had less than 5 years of impact factor data as of 2013 (not common), the average of existing years was used. Null values were used for the unmapped journals in the statistical calculations.

To estimate the prevalence of bioinformatics papers within MEDLINE using the same criteria as used for the analysis of the top 20 most cited papers, queries were constructed using Medical SubHeadings (MeSH). First, all queries were restricted to the years in question 1994–2013, reviews were excluded, and only papers with abstracts were included (to screen out news stories and editorials). MeSH terms are assigned to each abstract by curators, who also annotate the 'major topics' of a paper. The major topics chosen to search for bioinformatics papers were 'algorithms', 'software', 'bioinformatics' (synonymous with 'Computational Biology' in MeSH) and 'databases, factual'. On the most stringent setting, there would be no expansion of the major topics into their subtopics (e.g. 'Software' subtopics include video games, web browsers and word processors), Less stringent would be to allow expansion to subtopics, and least stringent would be to simply look for any of these annotations and their subcategories within any of the MeSH headings assigned to a paper. To calculate the Gini coefficients, citation counts were obtained from MEDLINE for all non-review articles for each journal for the entire period of analysis (1994–2013) and analyzed together to enable a calculation of their general, long-term trend.

2.2 Classification of publications

Classification of a paper as a bioinformatics paper was not always straightforward. For example, some report primarily on the software with minimal discussion on the underlying algorithms (e.g. Kumar *et al.*, 2001), while others describe statistical or mathematical methods/formulas without providing software (e.g. Cortes and Vapnik, 1995). Papers such as the initial Gene Ontology report (Ashburner *et al.*, 2000) were classified as a bioinformatics paper not because they provide analysis software, but because one goal of the GO was to create an accessible resource that enabled the computational analysis of scientific data. It could reasonably be argued that papers primarily about mathematical or statistical methods without software implementation, particularly those that are general and could be used in many fields, are not really bioinformatics papers. For example, papers that reported the development of general analysis methods such as support vector machines (Cortes and Vapnik, 1995) or a method to estimate the false discovery rate (Benjamini and Hochberg, 1995). These are methods that are widely implemented in bioinformatics software but were not designed specifically for biomedical applications. Thus, these types of papers were classified as 'methods' papers and considered both separately and as bioinformatics-related papers to see how the results were affected.

Additionally, distinctions are sometimes drawn between bioinformatics and computational biology, the latter being more oriented towards computational analysis of molecular objects in three or four dimensions. In this report, the term 'bioinformatics' will take the broader definition of the term and encompass all computational methods of biological data analysis.

2.3 Analysis of the highest impact papers in science

The start date of 1994 was chosen because it is approximately when the Internet Age began, during a time when personal computers became increasingly affordable and web browsers enabled non-programmers to more easily navigate the World Wide Web. The year 2013 was chosen as the end point because, even though statistics are available for 2014–2015, membership in the Top 20 for more recent papers is likely to be less stable and, therefore less informative. Analysis was limited to the top 20 papers partly to keep the problem tractable, but also because the number of papers follows an exponential curve and the 20th paper, very roughly, is the approximate point where many curves begin to inflect (Fig. 1) whereby papers on the left (most cited) end will be substantially more stable as members of the top 20 as time goes by than papers that are separated in their rank by fewer citations. The left side of this curve also represents papers that could be said to have had a disproportionate impact on science. For example, in Figure 1, the top 20 papers garnered 14.2% of the citations to the top 500 papers, despite being only 4% of the total.

3 Results

The 400 top papers, 20 per year for 20 years contained a total of 113 bioinformatics papers (28.25%). When counting methods papers as bioinformatics-related papers, there were 136 (34%). In more than half the years evaluated (55%), a bioinformatics program was the most cited paper (Table 1). This is striking because the relative fraction of bioinformatics papers within the literature as a whole is much smaller.

To get an approximate estimate of how many bioinformatics papers are in MEDLINE, under the same definitions as used here to evaluate the Top 20 most cited, three MeSH-based queries were constructed varying in their stringency (see methods). As a baseline, 47/50 (94%) of papers from journals with 'bioinformatics' in their name were evaluated and determined to be reporting the development of novel software or methods (the others were analysis papers). The MeSH-based queries were launched and, as a positive control, the number of papers in journals containing 'bioinformatics' in their name using the same query was evaluated. A sampling of 50 articles from each query was manually evaluated to estimate the number of false-positives (FP) returned (Supplementary Table 3). As shown in Table 2, the coverage of the least restrictive query (90%) is close to the baseline estimate of the fraction of bioinformatics papers that contain software/methods development (94%). Using this estimate, approximately 1.1% of the papers in MEDLINE report the development of bioinformatics methods or software. Thus, bioinformatics papers are approximately 26 to 31-fold over-represented among science's most cited papers, depending on whether methods papers are counted as bioinformatics.

A comparison of the journal impact factors (JIF) between the annually most highly cited bioinformatics papers and non-bioinformatics paper shows that despite their disproportionate impact on science, top 20 bioinformatics papers tend to be published in journals with a significantly lower average 5-year JIF (8.0) relative to the top non-bioinformatics paper (25.8, $P < 1.5 \times 10^{-24}$).

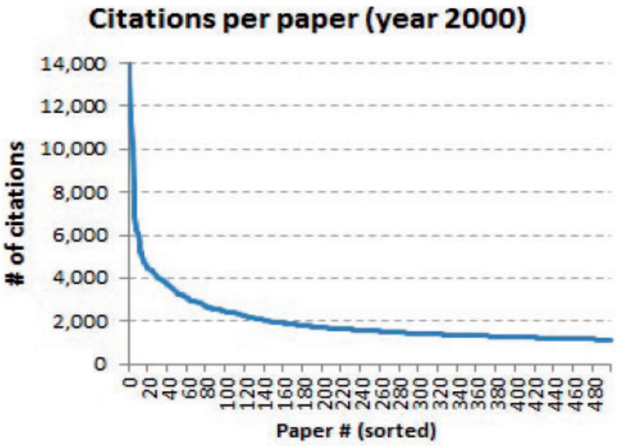

Fig. 1. The year 2000 as an example of the exponential-shaped curve typical of annual citation counts

Table 1. Most cited non-review articles from the approximate start of the Internet Age (∼1994) to 2013 according to the Institute for Scientific Information (ISI) Web of Knowledge

Most highly cited paper	Year published	Citations	# bioinf in Top 20	Avg bioinf JIF	Avg non-bioinf JIF
MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0	**2013**	**4531**	5	9.3	26.5
Observation of a new particle in the search for the Higgs boson	2012	3163	5	14.8	28.4
MEGA5: Molecular Evolutionary Genetics Analysis	**2011**	**19 098**	5	18.6	35.5
Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008	2010	5676	10	8.2	24.1
Systematic and integrative analysis of large gene lists using DAVID	**2009**	**6242**	7	7.5	23.3
A short history of SHELX	**2008**	**47 516**	8	10.2	29.5
MEGA4: Molecular evolutionary genetics analysis	**2007**	**20 470**	8	6.9	33.6
Induction of pluripotent stem cells from mouse embryonic cultures	2006	8503	5	10.8	23.9
Two-dimensional gas of massless Dirac fermions in graphene	2005	9091	5	5.8	25.5
Electric field effect in atomically thin carbon films	2004	20 395	11	5.4	30.5
MrBayes 3: Bayesian phylogenetic inference under mixed models	**2003**	**14 638**	11	8.6	21.1
The Cambridge Structural Database	**2002**	**8982**	6	4.1	26.4
Analysis of relative gene expression data using real-time quantitative PCR	2001	38 893	7	6.9	32.3
The Protein Data Bank	**2000**	**14 420**	4	6.8	23.1
From ultrasoft pseudopotentials to the projector augmented-wave method	1999	18 566	5	11.2	16.6
Crystallography & NMR system: A new software suite	**1998**	**15 269**	5	6.3	24.1
Gapped BLAST and PSI-BLAST	**1997**	**40 205**	10	5.8	32.8
Generalized gradient approximation made simple	1996	47 033	7	3.2	16.8
Controlling the false discovery rate	1995	21 224	7	3.2	27.1
CLUSTAL-W - improving sensitivity of multiple sequence alignment	**1994**	**42 995**	5	7.1	19.1

Citation data was compiled March 21, 2016 and data for all papers analyzed can be found in Supplementary Tables S1 and S2. Bioinformatics papers are **bolded**, and general methods papers frequently used in bioinformatics programs are *italicized*. Shown are the titles of the most cited papers each year (sometimes shortened to fit), the number of citations accrued at the time of this study (dataset citations from ISI's Data Citation Index not included), the number of bioinformatics (including methods) papers in the top 20 for each year, and the average JIF for the bioinformatics papers and non-bioinformatics papers for each year.

Table 2. Estimating the fraction of bioinformatics and methods papers in MEDLINE

	% bioinf found	# results	est. FP	% of lit	% of lit (-FP)
Most stringent	66%	84 809	20%	0.6%	0.48%
Less stringent	77%	145 409	42%	1.1%	0.64%
Least stringent	90%	319 150	54%	2.4%	1.10%

Shown are the fraction of all known bioinformatics papers during the query period returned by each query, the number of total results, the estimated false-positive rate judged by a sampling of 50 papers returned by the query (Supplementary Table S3), the percent of all papers published during this time the query returned, and the estimated number of bioinformatics papers by subtracting out the estimated false-positives.

Similarly, bioinformatics and methods papers combined had an average JIF of 7.7 versus 25.8 ($P < 4.5 \times 10^{-29}$).

3.1 Examining the robustness of the results

The top 20 selection criteria (no reviews, no dataset-based citations) were chosen to enable a comparison of literature-based citations to primary research, but to see how the results changed, the filters were removed. Supplementary Table S2 shows the results—the number of bioinformatics papers with the most citations dropped from 11 to 9, and the number of bioinformatics papers within the 400 total dropped to 80 (103 if including methods papers). The JIF disparity was still significant, and bioinformatics papers were still disproportionately enriched, from 18 to 23-fold.

The same data was also examined using the 2013 (2-year) JIF, but the results were not significantly changed (supplementary Tables S1 and S2).

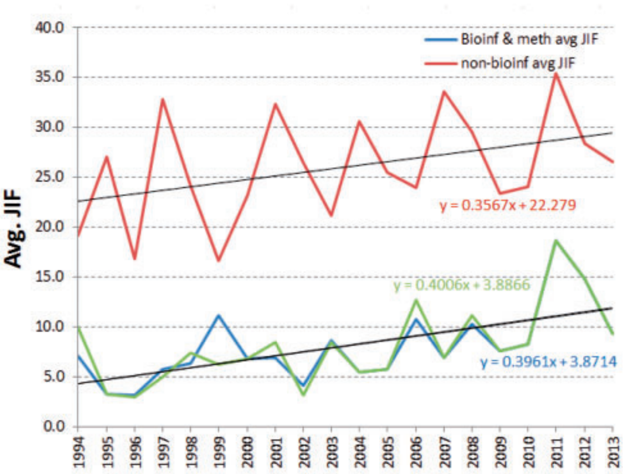

Fig. 2. Average journal impact factor (JIF) over the past 20 years for bioinformatics papers (green), bioinformatics and methods papers (blue) and for non-bioinformatics papers (red). Linear trend line slopes also shown

3.2 Is the journal impact factor gap narrowing?

It is reasonable to consider that, because bioinformatics was a relatively nascent field during much of the period analyzed, that perhaps it would simply take time for journals to recognize that bioinformatics approaches are consistently among the higher-impact papers each year and thus be incentivized to consider more such papers.

Plotting the average top 20 JIF for bioinformatics and non-bioinformatics papers over the two decades analyzed showed the top 20 non-bioinformatics paper JIFs were increasing at a similar rate (slope = 0.36) than bioinformatics papers (slope = 0.40) (Fig. 2). Although it might be argued the gap is slowly closing, there is quite a bit of year-to-year variability, making it difficult to draw strong conclusions.

3.3 What might account for the JIF-citation gap?

If bioinformatics journals publish a disproportionate number of the most highly cited papers, then it's reasonable to ask why their JIFs aren't higher. Although the number of citations per paper in journals normally follows a power law distribution, one possibility might be that there is an even greater skew among bioinformatics papers. To examine this, the Gini coefficient was calculated for 12 of the journals that most frequently published a top 20 most-cited paper between 1994 and 2013. The Gini coefficient is more commonly used as a metric of income inequality, but can also be used to represent citation inequality (Nuti *et al.*, 2015). If every paper published in a journal is cited equally, then the Gini coefficient would be 0. If all the citations went to one paper, it would be 1.0. Citations (see methods) were compiled for all non-review journal articles for the period 1994–2013 and analyzed as a whole (i.e. not by year). Table 3 shows that journals focusing on bioinformatics methods tend to have higher Gini coefficients.

4 Discussion

The higher Gini coefficients for bioinformatics journals suggest that development of novel bioinformatics resources may be somewhat 'hit or miss'. That is, some approaches become widely adopted and produce a disproportionate number of extremely highly cited papers, while most are not widely adopted and, at least relative to some of the other journals analyzed here, there are not as many papers that fall between the two extremes. This may also provide a potential explanation for the JIF gap: Methods developed to solve biological problems, whether novel ones or significant improvements on prior methods, may be technically sound but if it is difficult to know *a priori* which ones will be widely adopted or sorely needed, then it would be understandable for that to diminish enthusiasm.

There is a degree of subjectivity when classifying papers strictly into 'bioinformatics' versus 'non-bioinformatics', as some cases are not clear (e.g. should a minor permutation of an existing method be counted as a new method?). But the magnitude of the differences found suggests that the main conclusions are not likely to be altered by a few differences of opinion regarding paper classification.

Table 3. Analysis of Gini coefficients from 12 of the journals with highly cited papers

Journal	# papers	Gini coeff
J Virol	26 351	0.49
Blood	23 680	0.52
Cell	7186	0.52
PNAS	62 046	0.54
Mol Biol Evol	4031	0.63
Genome Biol	2028	0.67
BMC Bioinformatics	**5862**	**0.68**
N Engl J Med	6916	0.68
Nature	17 213	0.68
Nucleic Acids Res	19 828	0.69
Lancet	8 802	0.70
Science	18 756	0.73
Bioinformatics	**8369**	**0.76**
Acta Crystallogr D	**4856**	**0.91**
J Comput Chem	**2845**	**0.94**

BMC Bioinformatics did not have any top 20 papers, but was added for additional general perspective on bioinformatics journals (**bolded**). Similarly, *J Virol* and *Blood* were also added as representative of long-standing 'mainstream' biomedical journals.

4.1 Alternative metrics to estimate bioinformatics impact

If JIF is not a good estimate of the relative scientific impact of a bioinformatics program, then it is worth considering alternative metrics of impact, or 'altmetrics' as they have come to be known (Piwowar, 2013). There are a large number of possibilities in general (Holbrook *et al.*, 2013), but the primary issues would be which ones would best help normalize recognition in the identified problem areas for bioinformatics.

At least three challenges have been identified in the recognition of bioinformatics-related contributions to science. The first is career advancement for bioinformaticists that make most of their contributions as collaborators (Chang, 2015). It seems reasonable that, as long as the H-index metric is not weighted to penalize author position on the byline as has been proposed (Zhang, 2009), that this would reward collaborative ('team') efforts. However, the main point that not all author positions are equal is still valid, particularly in light of the fact that the average number of authors per paper in MEDLINE has been steadily rising (Hennessey *et al.*, 2014), and this would not address that.

Second, for bioinformatics research that does not result in a publication, the general solution would be to quantify how useful the program was to others. Recently, for example, Depsy.org was launched to quantify the impact of the 'software that powers science' (Singh Chawla, 2016). It combines full-text mining for mention of software packages with software archive statistics on the reuse of code and number of downloads.

Third, for researchers publishing their bioinformatics solutions to solve biomedical problems, this study has highlighted the gap between the retrospective impact of highly influential papers and their initial perceived importance. One alternative metric might be to use the number of citations instead of JIF, as citations are both quantitative and already well accepted. The downside is that citations take time, whereas JIF is known immediately. Plus, the average scientist can more easily estimate the relative significance of an impact factor number than a raw citation count.

4.2 Investing superstar updates in bioinformatics journals

Table 1 shows that three of the major updates to MEGA (Molecular Evolutionary Genetics Analysis), a phylogenetics program, were the most cited paper in the year they were published. Bioinformatics, as a field, is somewhat unique in that the product of bioinformatics research can continually evolve. Yet, not unlike experimental methods papers, such as methods to analyze quantitative PCR (qPCR) data, which also appeared three times in the top 20. Authors of 'superstar' bioinformatics programs (defined loosely here as those that appear in the top 20 most cited papers of any year) may be understandably tempted to publish in journals with the highest JIF. But if they consider that they also have many other bioinformatics projects that have not reached superstar status, then they would also recognize at least two things: First, the number of citations to their program alone and the fact it was among the most cited papers for that year is a reportable accomplishment on any *curriculum vitae*, regardless of the journal it appeared in, and likely sets it apart from the majority of papers in any journal anyway. Second, by publishing it in a bioinformatics-related journal historically associated with other superstar papers (e.g. Table 4), they are essentially

Table 4. Journals with the most bioinformatics or methods papers within the 20 most cited papers by year for the period 1994–2013

Journal	# Top 20 bioinf/ methods papers	Total # of top 20 papers '94-'13	% that were bioinf papers
Bioinformatics	15	15	100%
Acta Crystallographica Section D	9	9	100%
Molecular Biology and Evolution	8	8	100%
Genome Biology	5	5	100%
Systematic Biology	4	4	100%
Journal of Applied crystallography	3	3	100%
Journal of Molecular Biology	3	3	100%
Machine Learning	3	3	100%
Nucleic Acids Research	11	10	91%
J Computational Chemistry	5	4	80%
Physical Review B	5	4	80%
Genome Research	4	3	75%

Only journals with at least 3 papers, 75% or more of which were bioinformatics-related are shown.

boosting the impact of all future papers they and their colleagues publish there by association. So the motives for preferring traditional bioinformatics-oriented journals as a venue to publish superstar update reports need not be purely altruistic.

5 Conclusion

This study highlights the disproportionate impact bioinformatics has had on science in the past two decades. The Gini coefficient analysis of citations by journal suggests a greater gap between citation-rich and citation-poor bioinformatics papers, which suggests that it may be worth examining in greater depth why some of the less cited methods fared so poorly—were they simply not needed or was the right audience simply not aware of their existence? Are researchers reluctant to adopt new methods, even if better, if a solution already exists? It suggests bioinformatics, as a field is somewhat 'hit or miss' or high-risk/high-reward, relative to other fields.

As researchers, bioinformaticians face challenges in being recognized for their work that are somewhat unique to their field. It's tempting to suggest a change to the system is in order, but outside of bioinformatics there seems little incentive to do so. The use of altmetrics may be one immediate solution, but that presumes that others will officially recognize and accept their use. Another solution is for authors of superstar papers to invest in bioinformatics as a field by publishing substantial updates and improvements to their programs in bioinformatics-related journals. Awareness of the underestimated impact of bioinformatics is only part of the problem, the other part is finding effective solutions.

Funding

We would like to thank the National Science Foundation (grant # ACI-1345426) for support.

Conflict of Interest: none declared.

References

Aho,V.A. *et al*. (1976) Bounds on the complexity of the longest common subsequences problem. *J. ACM*, **23**, 1–12.

Althouse,B.M. *et al*. (2009) Differences in impact factor across fields and over time. *JASIST*, **60**, 27–34.

Anderson,J.S. *et al*. (1984) *Nucleotide Sequences 1984: A Compilation from the GenBank and EMBL Data Libraries: A Special Supplement to Nucleic Acids Research, Part 2*. IRL, University of California.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.

Beynon,R.J. (1985) CABIOS Editorial. *Comput. Appl. Biosci.: CABIOS*, **1**, 1.

Chang,J. (2015) Core services: reward bioinformaticians. *Nature*, **520**, 151–152.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn*., **20**, 273–297.

Doolittle,R.F. (1981) Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149–159.

Hennessey,J. *et al*. (2014) Trends in the production of scientific data analysis resources. *BMC Bioinformatics*, **15**, S7.

Holbrook,J.B. *et al*. (2013) Research impact: we need negative metrics too. *Nature*, **497**, 439-439.

Howison,J. and Bullard,J. (2015) Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature, *JASIST*, doi:10.1002/asi.23538.

Koch,R.E. (1971) The influence of neighboring base pairs upon base-pair substitution mutation rates. *Proc. Natl. Acad. Sci. USA*, **68**, 773–776.

Kumar,S. *et al*. (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, **17**, 1244–1245.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*., **48**, 443–453.

Nuti,S.V. *et al*. (2015) Association between journal citation distribution and impact factor. A novel application of the Gini coefficient. *J. Am. Coll. Cardiol*., **65**, 1711–1712.

Ohta,T. and Kimura,M. (1971) Functional organization of genetic material as a product of molecular evolution. *Nature*, **233**, 118–119.

Ouzounis,C.A. (2012) Rise and demise of bioinformatics? Promise and progress. *PLoS Comput. Biol*., **8**, e1002487.

Ouzounis,C.A. and Valencia,A. (2003) Early bioinformatics: the birth of a discipline – a personal view, *Bioinformatics*. **19**, 2176–2190.

Perez-Iratxeta,C. *et al*. (2007) Evolving research trends in bioinformatics. *Brief. Bioinf*., **8**, 88–95.

Piwowar,H. (2013) Altmetrics: value all research products. *Nature*, **493**, 159.

Rung,J. and Brazma,A. (2013) Reuse of public genome-wide gene expression data. *Nat. Rev. Genet*., **14**, 89–99.

Sander,C. (2001) Bioinformatics Editorial. *Bioinformatics*, **17**, 1–2.

Schwarzenbach,D. *et al*. (2010) SHELX makes an impact. *Acta Crystall. Section A: Found. Cryst*., **66**, 631.

Singh Chawla,D. (2016) The unsung heroes of scientific software. *Nature*, **529**, 115–116.

Van Noorden,R. *et al*. (2014) The top 100 papers. *Nature*, **514**, 550–553.

Wren,J.D. *et al*. (2007) The write position. A survey of perceived contributions to papers based on byline position and number of authors. *EMBO Rep*., **8**, 988–991.

Zhang,C.T. (2009) A proposal for calculating weighted citations based on author rank. *EMBO Rep*., **10**, 416–417.