Bioinformatics, 36(9), 2020, 2958–2962
doi: 10.1093/bioinformatics/btaa071
Advance Access Publication Date: 4 February 2020
Letter to the Editor

Data and text mining

An examination of citation-based impact of the computational biology conferences

Jayvardan S. Naidu[1], Justin D. Delano[1], Scott Mathews[2] and Predrag Radivojac [1,*]

[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA and [2]Department of Computer Science, Indiana University Bloomington, IN 47408, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Contact: predrag@northeastern.edu

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

Received on June 30, 2019; revised on December 18, 2019; editorial decision on January 19, 2020; accepted on January 27, 2020

1 Introduction

The impact of scientific publication is a multifaceted concept that is often measured through some form of citation analysis (van Wesel, 2016). The appropriateness of these analyses has been continually debated (Van Noorden, 2010) with many measures proposed to establish the prestige of individuals, departments, journals and conferences, both within and across disciplines (Bollen *et al.*, 2009; Garfield, 1955; Gross and Gross, 1927; Hirsch, 2005; Kaur *et al.*, 2013; Radicchi and Castellano, 2012; Schreiber, 2008; Vucetic *et al.*, 2018). Though not fully understood, the implications of such characterizations are believed to be wide-ranging, from summarizing an individual's academic performance to informing science policy (Fortunato *et al.*, 2018).

The difficulties with evaluating scientific impact generally arise from the inability of single-number summaries to capture all intellectual aspects of published work as well as disciplinary idiosyncrasies such as publishing norms and citation practices (Radicchi and Castellano, 2012; Van Noorden, 2010; Wang *et al.*, 2017). Other considerations involve the proper treatment of interdisciplinary work and multi-disciplinary individuals, team science and system abuse (Lopez-Cozar *et al.*, 2012), as well as the incorporation of the evolving nature of these factors over time (Milojević, 2014). Although various measures have been explicitly designed to address this problem and achieve universality, it remains most meaningful to limit such analyses to researchers in the same field and venues that are similar in scope.

The objective of this work was to investigate and summarize the citation-based impact of top conferences in the field of computational biology in terms of primary research published at those venues. We required that these venues had a multi-year tradition of soliciting full-paper and methodologically oriented manuscripts, with a submission deadline and predetermined review period, as well as the expectation that a large fraction of the accepted papers would be orally presented by one of the authors. Though within-field uniformity simplifies comparisons, the task is not straightforward due to the difficulties in extracting individual papers, incorporating journal extensions, tracking their citations over time and finding appropriate measures to summarize the results.

We looked at five major venues: the Intelligent Systems for Molecular Biology (ISMB), Pacific Symposium on Biocomputing (PSB), Research in Computational Molecular Biology (RECOMB), European Conference on Computational Biology (ECCB) and the ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB). These conferences have similar practices and overlapping communities, and thus desirable commonalities for our analysis. Our results suggest that all venues are influential, but in terms of individual summaries, we see trends supporting ISMB as the most impactful meeting.

2 Materials and methods

2.1 Datasets

The collection of papers and the citation data was performed manually. The lists of accepted papers for each conference were gathered from the conference website and cross-checked with the dblp database (https://dblp.uni-trier.de). After collecting the proceedings for each conference, each paper was run against Google Scholar's database (https://scholar.google.com) in order to collect the appropriate citation data. The data is available at https://github.com/predragradivojac/conferences.

Overall, our data contain 3707 papers, published between 1993 and 2017, written by more than 7000 individual authors at the rate of 3.7 authors per paper. Per conference, there are 1110 ISMB papers (3.6 authors per paper), including 332 jointly published with ECCB (3.7), 997 PSB papers (3.9), 773 RECOMB papers (3.5), 338 ECCB papers (3.9) and 489 BCB papers (3.6).

2.2 Conferences

ISMB was first held in 1993 and, up until 2000, its papers were released as proceedings issues under the Association for the Advancement of Artificial Intelligence (AAAI). From 2001 onwards, ISMB papers have been published in the *Bioinformatics* journal under the International Society for Computational Biology (ISCB). However, the *Bioinformatics* special issue usually includes a number of additional articles that are not part of ISMB proceedings. Therefore, we manually inspected all *Bioinformatics* issues

containing ISMB papers and removed articles that did not appear in the conference.

ISMB and ECCB are sometimes held jointly. The first such event happened in 2004. Additionally, since 2007, ISMB has been jointly organized with ECCB in odd-numbered years. There were six such events between 2007 and 2017. In our main analysis, joint ISMB-ECCB events were assigned to ISMB, as the more 'senior' event; however, a full breakdown and detailed analyses, when ISMB-ECCB was treated as a separate event, are shown in Supplementary Materials.

PSB started in 1996 and has not made significant changes in its publication model since its inception. Collection of manuscripts and citations was therefore straightforward. All PSB papers are exclusive to the conference with no exceptions, and the potential journal follow-up publications were considered external to the conference. For the purpose of collecting all research articles published in PSB, we excluded all session introduction and all workshop papers.

RECOMB started in 1997 and appears to have made several changes in its publication model, which resulted in some exception handling. Though RECOMB traditionally publishes conference proceedings, some articles have journal extensions. Additionally, in early years some articles might not have been peer-reviewed (abstracts), whereas in later years proceedings articles might only be published as abstracts with the actual paper occurring as a journal publication elsewhere or as an article on some of the preprint servers such as arXiv or bioRxiv.

Citations had to be properly combined to ensure that the citation count for each paper was accurate. We followed a certain set of rules to ensure consistency of citation combination. Considering a given full proceedings paper or an extended abstract, if there existed a journal publication with similar content (graphics, identical paragraphs, etc.), their citations were summed, except in about 7% of cases where Google Scholar already combined the articles/citations. If the authors on the journal publication were a proper superset of the authors in the RECOMB proceedings article, citations were not combined; however, if they were a subset, then the citations were combined. The rationale was that journal follow-ups could be substantially modified and have a very different focus.

The aggregation of citations for journal and conference articles had the potential to overcount the real impact based on the fact that some publications might cite both articles, including the possibility that the journal paper cited the conference paper it was an extension of. We have manually inspected a random sample of 10 such articles and concluded that the impact of double citations was small, estimated at under 5% of the total per paper combined citation count.

ECCB started in 2002 with its proceedings papers published in special issues of the *Bioinformatics* journal. Here, we encountered the same issues that we ran into while processing ISMB papers. Moreover, in 2002 and 2003, the special issue of ECCB also included publications that were not full research papers. Those papers were excluded from our analysis.

BCB started in 2010 and is published under ACM as a proceedings. We only included 'regular papers' and disregarded 'short papers'. Collecting citation data for the BCB papers was relatively simple. Occasionally, a paper would also be posted in a journal. It was therefore important to make sure the citation count for the given paper included citations that were intended for the corresponding journal article. We followed the same rules as for the RECOMB papers.

2.3 Conference similarity

We calculated the similarity of conferences by looking at the overlap of authors who published in these venues. We encoded each conference either as a set of unique authors who published there over the years in consideration or as a bag-of-words representation, where we recorded the number of times an author published in a particular venue. We were not able to disambiguate names, and thus, the authors were represented as concatenations of strings containing the first initial and the last name. Distinct individuals with identical string representations resulted in 'combined' authors. Similarly, individuals with more than a single name in the analyzed time span would ultimately contribute to 'split' authors. Jaccard (1901) distance was used to measure set similarity (one set of authors for each conference) and normalized Yang–Clark distance (Clark and Radivojac, 2013; Yang *et al.*, 2019) with $p = 2$ was used to measure the similarity of integer-valued vectors (one vector of publication counts per author for each conference).

More specifically, let X and Y be sets of authors publishing in conferences c_X and c_Y, respectively. Let also $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$ be bag-of-words representations of authors publishing in c_X and c_Y, respectively, with x_i and y_i being the numbers of papers an author indexed by i published in a given period of time in c_X and c_Y. The Jaccard distance metric between the two conferences was calculated as

$$d_J(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \setminus Y| + |Y \setminus X|}{|X \cup Y|},$$

where $|\cdot|$ is the cardinality of the set. Similarly, the normalized Yang–Clark distance metric of order $p \geq 1$ was calculated as

$$d_{YC}(x, y) = \frac{\left(\left(\sum_{x_i \geq y_i}(x_i - y_i)\right)^p + \left(\sum_{y_i > x_i}(y_i - x_i)\right)^p\right)^{\frac{1}{p}}}{\sum \max\{|x_i|, |y_i|, |x_i - y_i|\}},$$

where $|\cdot|$ is the absolute value function. When $p = 1$, this distance is the real-valued equivalent of Jaccard distance, that is, when inputs are real-valued vectors instead of sets. We note that all distances between venues were calculated only over the years in which both conferences were held. Once distances were computed for each pair of entries, the conferences were clustered using hierarchical clustering (Tan *et al.*, 2006).

2.4 Statistical significance of differential conference impact

For the purposes of this analysis, the conference c_X is considered more impactful than conference c_Y if a paper (average, median or median of top 10 papers, as we will elaborate on later) published in any given year at c_X is expected to collect more citations in subsequent years than a paper published at c_Y. To quantify the statistical surprise for the observed citation differences between the two conferences, we performed one-sided binomial tests, where the null hypothesis was that c_X and c_Y were equally impactful and the alternative hypothesis was that c_X was more impactful than c_Y. Each comparison was performed over K years in which both c_X and c_Y were held, and the P-value (P) was calculated as

$$P = \sum_{i=k_0}^{K} \binom{K}{i} \alpha^i (1 - \alpha)^{K-i},$$

where k_0 is the number of years where we observed that c_X had more citations per paper than c_Y. The parameter α was set to $\frac{1}{2}$ to indicate an equal chance of a win under the null model. When analyzing the performance of the median paper, we also encountered ties. The first occurrence of a tie was counted as a win for the event with fewer wins and the subsequent ties were awarded in an alternating fashion.

We observe that the P-values depend on the number of comparisons, and thus a longer history allows us to make stronger assertions about the relative impact of the conferences. Each year was considered to be an independent event.

3 Results

3.1 Relative impact of conferences

To provide comparisons of computational biology conferences, we manually collected publications of all proceedings papers and associated journal papers from five venues: ISMB, PSB, RECOMB, ECCB and BCB. For each paper published in each conference, we collected its number of citations according to Google Scholar, concluding with the end of 2017.

Given a paper published in year i, there are two types of impact considered: (i) multi-year impact, where the citations were collected from years $i+1$, $i+2$, through 2017 and (ii) 2-year impact, where the citations were collected in years $i+1$ and $i+2$. We opted to exclude year i from the calculations because the conference publication times range from January (PSB) to September (ECCB, BCB) in any given year, which could have significant influence in some comparisons. For each conference, we then looked at the (i) citations per year per published paper, (ii) citations per year of the median paper and (iii) citations per year for the median of the top 10 cited papers. These schemes provide slightly different views on the impact of these conferences.

Figure 1 shows the multi-year citation-based impact, divided in three panels representing average paper, median paper and median of the top 10 cited papers. We selected this as our main measure in order to measure longer-term impact of published work. However, the relative performance between any papers published in different years is not as meaningful because the papers were available to the public over different time intervals. The 2-year performance comparisons, shown in Supplementary Materials, give similar results. Identical analyses for the case when ISMB-ECCB was considered as a separate conference are also provided in Supplementary Materials.

Detailed head-to-head comparisons are shown in Table 1. Based on the average, median and top 10 median citations, ISMB performs favorably against all other events; e.g., it scores in wins vs. ties vs. losses as follows: (21:0:0, 21:0:0, 19:0:2) over PSB, (14:0:6, 15:1:4, 16:0:4) over RECOMB, (8:0:1, 7:1:1, 8:0:1) over ECCB, and (7:0:0, 7:0:0, 7:0:0) over BCB. ECCB scores (5:0:4, 4:2:3, 5:0:4) over RECOMB, (8:0:1, 8:0:1, 8:0:1) over PSB, and (4:0:0, 4:0:0, 4:0:0) over BCB. RECOMB scores (19:0:1, 18:0:2, 16:0:4) over PSB, and (7:0:0, 7:0:0, 7:0:0) over BCB. Finally, PSB scores (7:0:0, 7:0:0, 6:0:1) over BCB. We conclude that ISMB has the most citations and most pairwise wins, followed by ECCB, RECOMB, PSB and BCB. The 2-year head-to-head comparisons are shown in Supplementary Materials. As before, the same analyses for the case when ISMB-ECCB was considered as a separate conference are also provided in Supplementary Materials. These results suggest that ISMB statistics are slightly improved when ISMB-ECCB results are included in ISMB.

3.2 Similarity of conferences

We calculated the similarity between each pair of conferences (c_X, c_Y) based on the authors publishing in these venues. The similarity for each pair was calculated only over the years where both conferences were held, using set distances and vector-space distances described in Section 2. The single-linkage clustering for the five venues, depicted in Figure 2, shows the highest similarity between ISMB and RECOMB, followed by PSB, ECCB and BCB. The clustering was identical when the single-linkage similarity between groups of objects was replaced by either average similarity or Ward's method for both distance functions (Tan *et al.*, 2006). The complete-linkage, on the other hand, swapped the order of ECCB and BCB (Jaccard distance) or swapped the order of ECCB and PSB (Yang–Clark distance). Given that the fluctuations are minor, we believe that our main finding is relatively unaffected by the type of clustering and the distance measure.

4 Discussion

Scientific conferences and related scholarly events occupy a unique space in the scientific enterprise as they reflect both research and social aspects within a discipline (Francisco *et al.*, 2011; Jeong *et al.*, 2009). In assessing their influence, we took a restricted approach and summarized the impact of primary research presented in five major conferences in bioinformatics and computational biology. We manually collected the data for each original research paper published in each of the venues and provided comparative evaluation based on several criteria. Overall, we quantified the citation-based impact of each conference and summarized trends and differences

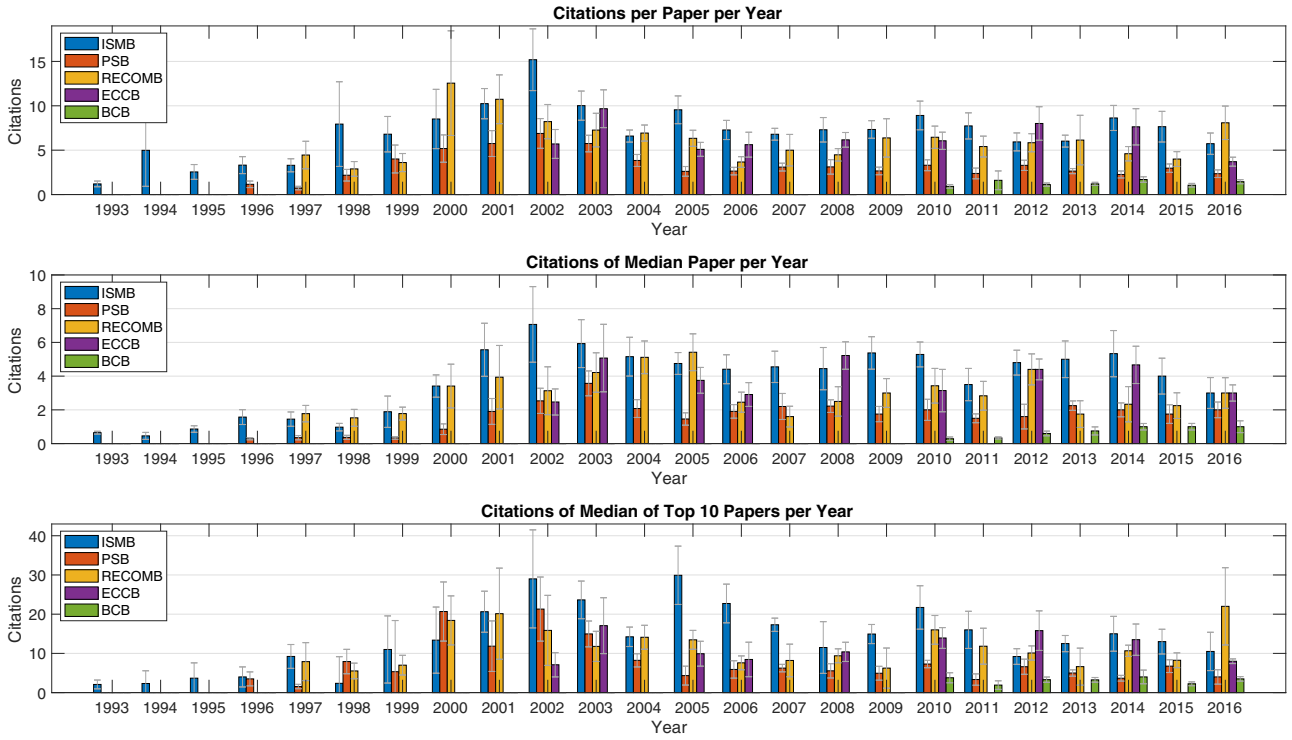

Fig. 1. Bar plots summarizing the number of citations per year for papers published in each conference. The number of citations for any paper published in year i was counted as the sum of all its citations from year $i+1$ up until end of 2017, divided by the number of years used to sum citations. Top panel shows the average number of citations per year for papers published in year i; middle panel shows the number of citations per year for the median paper published in year i; bottom panel shows the number of citations per year for the median of the top 10 cited papers published in year i. Standard errors were obtained by bootstrapping, with 1000 iterations, papers published in each conference and each year

Table 1 Multi-year head-to-head performance comparisons according to the average number of citations per paper per year, the average number of citations of the median paper and the average number of citations of the median of the top 10 papers

	ISMB				PSB				RECOMB				ECCB				BCB			
	W	T	L	*P*-value	W	T	L	*P*-value	W	T	L	*P*-value	W	T	L	*P*-value	W	T	L	*P*-value
Average																				
ISMB	—	—	—	—	21	0	0	$4.8 \cdot 10^{-7}$	14	0	6	$5.8 \cdot 10^{-2}$	8	0	1	$2.0 \cdot 10^{-2}$	7	0	0	$7.8 \cdot 10^{-3}$
PSB	0	0	21	1.0	—	—	—	—	1	0	19	1.0	1	0	8	1.0	7	0	0	$7.8 \cdot 10^{-3}$
RECOMB	6	0	14	$9.8 \cdot 10^{-1}$	19	0	1	$2.0 \cdot 10^{-5}$	—	—	—	—	4	0	5	$7.5 \cdot 10^{-1}$	7	0	0	$7.8 \cdot 10^{-3}$
ECCB	1	0	8	1.0	8	0	1	$2.0 \cdot 10^{-2}$	5	0	4	$5.0 \cdot 10^{-1}$	—	—	—	—	4	0	0	$6.3 \cdot 10^{-2}$
BCB	0	0	7	1.0	0	0	7	1.0	0	0	7	1.0	0	0	4	1.0	—	—	—	—
Median																				
ISMB	—	—	—	—	21	0	0	$4.8 \cdot 10^{-7}$	15	1	4	$2.1 \cdot 10^{-2}$	7	1	1	$9.0 \cdot 10^{-2}$	7	0	0	$7.8 \cdot 10^{-3}$
PSB	0	0	21	1.0	—	—	—	—	2	0	18	1.0	1	0	8	1.0	7	0	0	$7.8 \cdot 10^{-3}$
RECOMB	4	1	15	$9.9 \cdot 10^{-1}$	18	0	2	$2.0 \cdot 10^{-4}$	—	—	—	—	3	2	4	$7.5 \cdot 10^{-1}$	7	0	0	$7.8 \cdot 10^{-3}$
ECCB	1	1	7	$9.8 \cdot 10^{-1}$	8	0	1	$2.0 \cdot 10^{-2}$	4	2	3	$5.0 \cdot 10^{-1}$	—	—	—	—	4	0	0	$6.3 \cdot 10^{-2}$
BCB	0	0	7	1.0	0	0	7	1.0	0	0	7	1.0	0	0	4	1.0	—	—	—	—
Median of top 10																				
ISMB	—	—	—	—	19	0	2	$1.1 \cdot 10^{-4}$	16	0	4	$6.0 \cdot 10^{-3}$	8	0	1	$2.0 \cdot 10^{-2}$	7	0	0	$7.8 \cdot 10^{-3}$
PSB	2	0	19	1.0	—	—	—	—	4	0	16	1.0	1	0	8	1.0	6	0	1	$6.3 \cdot 10^{-2}$
RECOMB	4	0	16	1.0	16	0	4	$6.0 \cdot 10^{-3}$	—	—	—	—	4	0	5	$7.5 \cdot 10^{-1}$	7	0	0	$7.8 \cdot 10^{-3}$
ECCB	1	0	8	1.0	8	0	1	$2.0 \cdot 10^{-2}$	5	0	4	$5.0 \cdot 10^{-1}$	—	—	—	—	4	0	0	$6.3 \cdot 10^{-2}$
BCB	0	0	7	1.0	1	0	6	$9.9 \cdot 10^{-1}$	0	0	7	1.0	0	0	4	1.0	—	—	—	—

Wins (W), ties (T) and losses (L) were computed as the number of years a conference in row *i* had a better citation performance than the conference in column *j*. *P*-values were calculated as one-sided binomial tests that conference in row *i* outperforms the conference in column *j*, as described in Section 2.

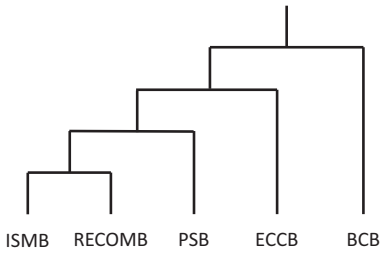

Fig. 2. Similarity of conferences calculated using the overlap of authors over the years each individual pair was held. Hierarchical clustering was used based on the Jaccard and Yang–Clark distances. The same dendrogram was obtained for both distance measures

among the venues. Our results suggest that the original research presented in all conferences is influential, with ISMB being more impactful than ECCB and RECOMB, that themselves have comparable statistics. Despite some outstanding years and strong top papers, PSB ranked next, and ahead of BCB.

Comparisons of scientific work are difficult and any attempt of comparative evaluation is undoubtedly limited. First, citations are unlikely to directly measure intellectual contributions, and thus might provide a distorted view of the quality of science. Second, there might exist subtle differences among conferences, such as subcommunities with different citation practices that could skew our conclusions. Third, during our data collection process we had to make several types of decisions on how to count publications, given different publication models adopted by each of the conferences, or disambiguate authors. While we believe our decisions were reasonable, this may not be universally agreed upon. Finally, we have omitted other conferences; e.g., Workshop on Algorithms in Bioinformatics (WABI), IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE International Conference on Bioinformatics and Biomedicine (BIBM), International Workshop on Data Mining in Bioinformatics (BIOKDD), that could have comparable impact to the ones analyzed herein.

On the other hand, citations have been shown to strongly correlate with peer-assessments of quality (Vucetic *et al.*, 2018) and when interpreted properly could be a useful metric. We further believe that the data collected in this work could be used to expand the analysis beyond publication venues and include individual or institutional statistics. Such analyses were beyond the scope of this report. Ultimately, these data could be useful in selecting program committees of future conferences based on the demonstrated stature and impact in the field associated with each author.

Conflict of Interest: Predrag Radivojac is an active Board of Directors member of the International Society for Computational Biology (ISCB) and a member of the Association of Computing Machinery (ACM). He has served as a Proceedings Chair (2018, 2019) and Area Chair (2006, 2007) of ISMB and ISMB-ECCB, a Session Chair of PSB (2006–2009, 2017 and 2019–2020), and a Vice Chair (2011) of BCB. He has been a Program Committee member for all studied conferences.

Financial Support: none declared.

References

Bollen,J. *et al.* (2009) A principal component analysis of 39 scientific impact measures. *PLoS One*, **4**, e6022.

Clark,W.T. and Radivojac,P. (2013) Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, **29**, i53–i61.

Dong,Y. *et al.* (2015) Will this paper increase your h-index? Scientific impact prediction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015*. ACM, pp. 149–158.

Fortunato,S. *et al.* (2018) Science of science. *Science*, **359**, eaao0185.

Francisco,M. *et al.* (2011) Conference models to bridge micro and macro studies of science. *J. Artif. Soc. Soc. Simul.*, **14**, 13.

Garfield,E. (1955) Citation indexes for science: a new dimension in documentation through association of ideas. *Science*, **122**, 108–111.

Gross,P.L.K. and Gross,E.M. (1927) College libraries and chemical education. *Science*, **66**, 385–389.

Hirsch,J.E. (2005) An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA*, **102**, 16569–16572.

Jaccard,P. (1901) Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin Del la Société Caudoise Des Sciences Naturelles*, **37**, 547–579.

Jeong,S. *et al*. (2009) Are you an invited speaker? A bibliometric analysis of elite groups for scholarly events in bioinformatics. *J. Assoc. Inf. Sci. Technol*., **60**, 1118–1131.

Kaur,J. *et al*. (2013) Universality of scholarly impact metrics. *J. Informetr*., **7**, 924–932.

Lopez-Cozar,E.D. *et al*. (2012) Manipulating Google Scholar citations and Google Scholar metrics: simple, easy and tempting. arXiv preprint arXiv: 1212.0638.

Milojević,S. (2014) Principles of scientific research team formation and evolution. *Proc. Natl. Acad. Sci. USA*, **111**, 3984–3989.

Radicchi,F. and Castellano,C. (2012) Testing the fairness of citation indicators for comparison across scientific domains: the case of fractional citation counts. *J. Informetr*., **6**, 121–130.

Schreiber,M. (2008) To share the fame in a fair way, *hm* modifies *h* for multi-authored manuscripts. *New J. Phys*., **10**, 040201.

Tan,P.N. *et al*. (2006) *Introduction to Data Mining*. Pearson, New York, NY.

Van Noorden,R. (2010) Metrics: a profusion of measures. *Nature*, **465**, 864–866.

van Wesel,M. (2016) Evaluation by citation: trends in publication behavior, evaluation criteria, and the strive for high impact publications. *Sci. Eng. Ethics*, **22**, 199–225.

Vucetic,S. *et al*. (2018) Peer assessment of CS doctoral programs shows strong correlation with faculty citations. *Commun. ACM*, **61**, 70–76.

Wang,J. *et al*. (2017) Bias against novelty in science: a cautionary tale for users of bibliometric indicators. *Res. Policy*, **46**, 1416–1436.

Yang,R. *et al*. (2019) A new class of metrics for learning on real-valued and structured data. *Data Min. Knowl. Disc*., **33**, 995–1016.