Trends in Gaming Indicators: On Failed Attempts at Deception and their Computerised Detection

Cyril Labbé

Trends in Gaming Indicators: On Failed Attempts at Deception and their Computerised Detection

Cyril Labbé

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
Cyril.Labbe@imag.fr

Abstract. Counting articles and citations, analyzing citations and co-authors graphs have become ways to assess researchers and institutions performance. Fairly enough, these measures are becoming targets for institutions and individual researchers thus triggering new behaviors. As a matter of fact, scientometrics and informetrics systems of all kinds have to separate the grain from the chaff. Among others, fields like information retrieval, network analysis and natural language processing may offer answers to deal with this kind of problems. Through several emblematic case studies (fake researcher, generated papers, paper mills), we show evidences of attempts to game indicators together with automatic ways to detect them (automatic detection of generated papers, errors detection).

1 Introduction

Several factors are substantially changing the way the scientific community shares its knowledge. On the one hand, technological developments have made the writing, publication and dissemination of documents quicker and easier. On the other hand, the pressure of individual and institutional evaluation is changing the publication process. This combination of factors has led to a rapid increase in scientific document production. In a sense, one could say that the global knowledge is growing ever faster than before. The presence of junk publications could be interpreted as a side effect of the 'publish or perish' paradigm.

Nevertheless, counting articles and citations, analyzing citations and co-authors graphs have become ways to assess researchers and institutions performance. Fairly enough, these measures are becoming targets for institutions and individual researchers thus triggering new behaviors. Several emblematic case studies (fake researcher, generated papers, paper mills) show evidences of attempts to game indicators. As a matter of fact, scientometrics and informetrics systems of all kinds have to sort out the publications that matters. Among others, fields like information retrieval, network analysis and natural language processing may offer answers to deal with this kind of problems.

The section 2 describes individual and institutional behavior that can be used to game metrics and ranking, some of them still in use. Section 3 exposes automatic ways to detect some of them.

2 Gaming indicators

Various kind of misconducts can be identified with regards to scientific publication. The interested reader may consult the Committee on Publication Ethics (COPE) catalog[1] which provides about 600 cases of such misconducts. Incentives for such practices may arise from very various and personal reasons. The following examples, intentionally leaving aside plagiarism, are chosen because they may be seen as clear examples of behaviors that are driven towards indicator manipulation.

Gaming University Ranking. According to the US News and World Report, in 2014, the King Abdulaziz University (KAU) in Jeddah, Saudi Arabia, was ranked 7th in the top ten universities in the mathematics area. Regarding the ARWU by subject field, in mathematics, KAU was ranked 51-75 in 2012 and reached the 6th position in 2015 (see figure 1). These results were achieved by literally buying publications and citations [1,2]. This is done by recruiting massively highly cited authors in a field, hiring them as *Distinguished Adjunct Professor* at KAU for them to list King Abdulaziz University as secondary affiliation. Lior Pastor reproduce[2] an e-mail stating the terms and conditions for joining the *International Affiliation Program* at KUA. These terms and conditions seem to include, for example, a per month salary of $6 000 and a mandatory visit of at least three weeks per year, KUA covering travel and living expenses for the visits. This illustrates how university rankings can be *manipulated*.

Hacking peer-review process. So-called *peer-review rings* are used to bypass real peer review and avoid rejection by gaining an easy and quick acceptation of submitted papers [3]. Such *peer-review rings* have been brought to light by the retractions of 64 articles in 10 Springer subscription journals[3].

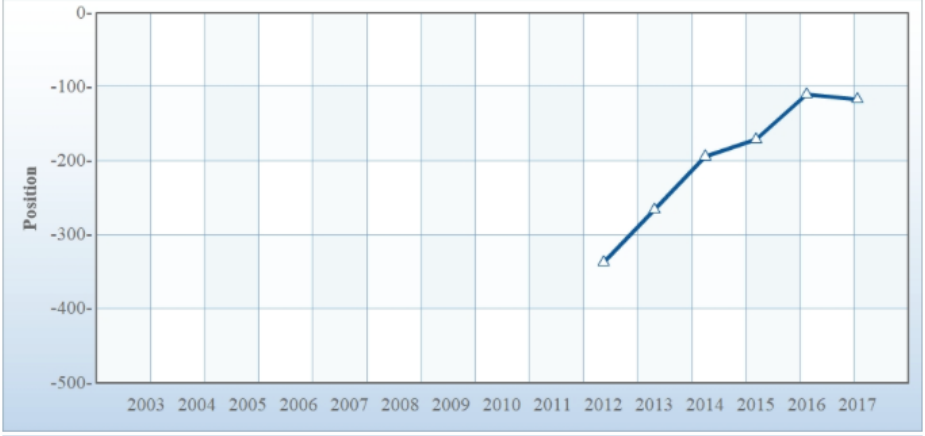
Recently *Retraction Watch* reported a case where the peer review process seems to have been used as a means to increase citations. Three papers were retracted from a journal because the proportion of citations to these papers were mostly from a single conference where an author of the retracted papers was chairing the conference.

[1] https://publicationethics.org/cases

[2] https://liorpachter.wordpress.com/2014/10/31/to-some-a-citation-is-worth-3-per-year

[3] http://www.springer.com/gp/about-springer/media/statements/retraction-of-articles-from-springer-journals/735218

Performance in Academic Ranking of World Universities



ARWU	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Institutional Ranking	/	/	/	/	/	/	/	/	/	301-400	201-300	151-200	151-200	101-150	101-150

Performance in Academic Ranking of World Universities by Broad Subject Fields

Broad Subject Fields	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
SCI	/	/	/	/	/	/	151-200	51-75	51-75	32
ENG	/	/	/	/	/	/	101-150	51-75	45	5
LIFE	/	/	/	/	/	/	/	/	/	/
MED	/	/	/	/	/	/	/	/	/	/
SOC	/	/	/	/	/	/	/	/	/	/

Performance in Academic Ranking of World Universities by Subject Fields

Subject Fields	2009	2010	2011	2012	2013	2014	2015
Mathematics	/	/	/	51-75	22	10	6
Physics	/	/	/	/	/	151-200	/
Chemistry	/	/	/	101-150	51-75	38	36
Computer Science	/	/	/	/	/	51-75	27
Economics/Business	/	/	/	/	/	/	/

Fig. 1: King Abdulaziz University by World University Rankings (`http://www.shanghairanking.com/World-University-Rankings/King-Abdulaziz-University.html`)

Ike Antkare the shooting star [4]. Without any regular publications in any conference proceedings, journal or other venue, for a few months, *Ike Antkare*[4] was ranked at the top of the academic charts, featuring a better score than Einstein and Turing. At this time, he was one of the most highly cited scientists of the modern world having 100 publications each of which were citing all others (including itself) together with an extra reference to another pseudo-document (referenced as Ike Antkare's PhD [5][5]) referencing only already indexed genuine documents.

Like a shooting star, Ike Antkare, was ranked directly in the 21st position of the most highly cited scientists (dixit scholarometer). In 2010, this score was less than Freud (1st position with a h-index of 183) but better than Einstein (36th position). Ike Antkare was at the top of the charts in rather good company getting well along with the Nobel price Paul Krugman, the inspiring Karl Marx and other famous names of his own field. Best of all, with regards to the h_m-index (which takes into account co-authorships to reward single-authored papers) Ike Antkare was in sixth position outclassing all scientists in his field (computer science).

Academic search engine optimization. A team of Spanish researchers Lopez-Cozar et al. reproduced a similar experiment [6] by making Google Scholar indexing fake citations to their own publications. This study shows the impact of such manipulation on their own h-index. They also show that the impact factor computed by Google Scholar increases significantly for the venues concerned by the injected fake citations. Logically it can be inferred that this is also true for labs and universities hosting these researchers. Genuine, border-line and un-recommended ways to increase the visibility of a particular work in Google Scholar have been studied by Beel et al. in [7,8]. This so-called *Academic search engine optimization* includes strategies ranging from making sure that the text can be properly extracted from PDF files and figures to the insertion of hidden references (white text over white background).

Fake papers make it through peer review. Automatically/generated fake scientific papers were spotted in several venues where they should not have been published, given the stringent process of selection they were supposed to have gone through. More than 100 SCIgen papers have purely and simply vanished from IEEE databases once they were exposed by Labbé and Labbé [9] and publicise by Van Noorden [10]. These papers were accepted in peer-reviewed conferences that sometimes claim an acceptance rate as strict as 28%. An example is the SSME conference once indexed by the Web of knowledge and Scopus. It was held in 2009 with 150 published papers. Among these 150 papers there were four SCIgen papers and one duplicate: two papers having exactly the same text but a different title. You have to think that these papers have been formally

[4] to be interpreted as *I can't care*

[5] This reference is not referencing any "real" scientific publication, but the document itself exists online. Details may be found in [4]

reviewed... presented to the conference audience of roughly 150 people... and discussed face to face, at least by a polite chair(wo)man. An investigation carried out by the journalist Shuyang Chen[6] shows that these papers were published mainly to fulfill the quantitative goals assigned to academics by the Chinese administration.

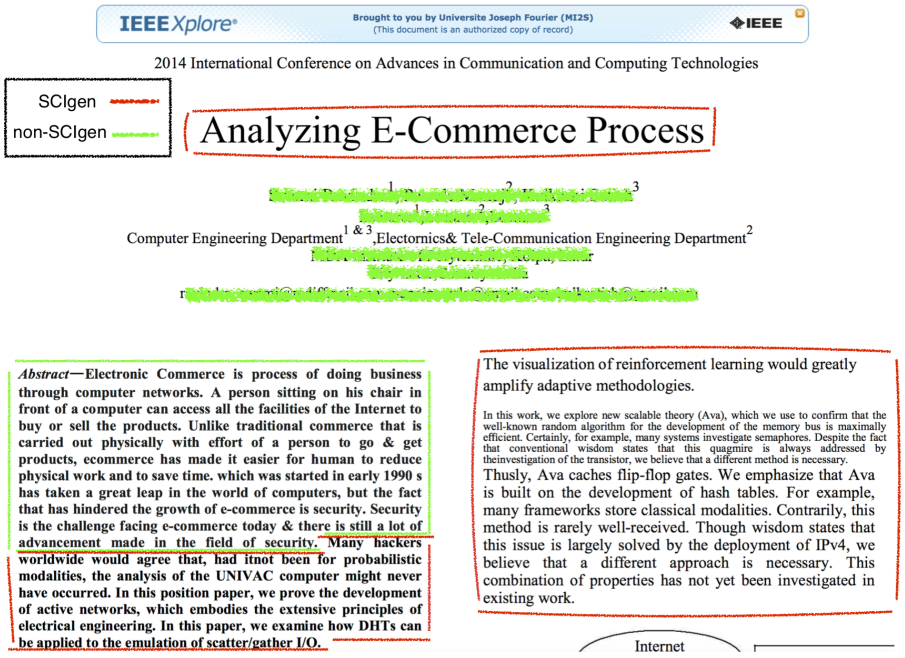


Fig. 2: Genuine and SCIgen text mixed up in a paper published in the 2014 International Conference on Advances in Communication and Computing Technologies (ICACACT): publisher IEEE.

The most recent example of such a paper is shown in figure 2. This paper [11] is itself a very interesting paper because it is a mix of SCIgen text intermingled with non-SCIgen text. It is also very interesting because authors are not from China which is the place where this kind of paper usually comes from. This paper remained unnoticed – and sold – for almost two years: the conference date is August 2014 and it was retracted in March 2016.

Paper mills and errors spreading. The use of paper mills and the possibility that "assisted" manuscripts may be produced on a large scale seems to be more and more evident [12]. Moreover, in growing cases the quality of the published

[6] `http://www.time-weekly.com/uploadfile/2014/0410/280.pdf` english translation available at `http://membres-lig.imag.fr/labbe/TimeWeekly.pdf`

results seems to be questionable. As example, IEEE is used to remove conferences from IEEE Xplore. An online form[7] can be used by authors of papers published in removed conferences if they wish a confirmation of their copyright ownership. According to the *conferences impacted* listed in this form more than 100 conferences are concerned.

As another evidence [13] reports preliminary evidence that education/biotechnology companies may be providing content pertaining to gene knockdown experiments in human cancer cell lines to researchers based in China, who then publish these results without disclosing their origin. This led to several retractions of published papers [14].

Conclusion. As the pressure to publish increases, scientific information systems – going from social networks to peer reviewed venues – are getting increasingly exposed to forged papers and papers containing errors. As a matter of fact, one can find them almost in every place where genuine scientific papers can be found. In this context automatic detection of such problematic publications becomes mandatory to ensure systems credit and renown.

3 Automatic detection of dubious behavior

Spotting dubious publications, dubious scientific results, non-relevant publications, citations or behavior is important to insure trust in science. The followings are examples of attempts to automatically detect some of these behaviors.

Fake paper detection Several methods have been developed to automatically identify SCIgen papers. For all of them, the first step is to extract the text from PDF files and then try to determine if this text is generated or not.

For example, Xiong and Huang [15] detect SCIgen paper by checking whether references, in the references section, are valid references. A reference is valid if it already exists in a trusted database. Following this approach, a paper with a large proportion of unidentified references will be suspected to be a SCIgen paper.

Lavoie and Krishnamoorthy in [16] an ad-hoc similarity measure between papers is defined aiming at extracting particular features of generated texts. In this measure the reference section plays a major role along with title and keywords. This is why this method failed to detect papers generated for the Ike Antkare experiment because of their very special references sections.

Dalkilic et al. [17] method is based on observed compression factor and a classifier. The goal in this study is more general than only detecting SCIgen papers. The idea is based on the fact that randomly generated texts (called inauthentic texts) do not have the same compression factor than non-random texts.

[7] `https://www.ieee.org/conferences_events/conferences/publishing/author_form.html`

Amancio [18] proposes a comparison of topological properties between natural and generated texts, and Williams and Giles in [19] studies the effectiveness of different measures to detect fake scientific papers.

Scientific information systems are so exposed to SCIgen threat, that even a premier open repository like ArXiv includes automated tests in order to detect possible fake papers. Ginsparg [20] method relies on characterizing the statistical distribution of a set of predefined stop-words. It seems that the method is quite effective and operative, as not a single SCigen paper was ever reported being "accepted" in ArXiv. This suggests that a well-managed open and non-peer reviewed system contains less gibberish than an expensive fee-based service.

Labbé's method [9] is based on inter-textual distance. For a text under consideration, the distances between the text and some previously known SCIgen are computed. When the SCIgen nearest neighbor is too close to the text under consideration then this latter is classified as a SCIgen text. A demonstration website for this method was set up and it soon started to be used quite heavily by publishers to make sure they will not accept SCIgen paper. Springer Nature funded the development of SciDetect an open-source software aiming at detecting all kind of known generators[8][21,22].

Citations Analysis. Citation analysis is also a means to detect attempts to manipulate indicators. For example, Bartneck and Servaas [23] investigates the possibility to detect h-index manipulation through the analysis of self-citations. On a similar topic, Herteliu et al. [24] shows that sometimes editors misbehavior may be detectable trough quantitative citation analysis. Fister et al. [25] also suggests that the analysis of *Citation Cartels* in citation networks could be of some help. Such kind of technics could, for sure, be used to detect a complete graph of citations such as the one used for the Ike Antkare experiments.

Errors detection. Errors within scientific publications contribute to research irreproducibility. To highlight errors or dubious publications, one can employ automatic approaches to check the statistical validity of presented values as done by Nuijten et al. [26].

Another approach is proposed by Labbé and Byrne [27] with the Seek & Blastn tools. This tool is a semi-automated tool that checks the claimed use of nucleotide sequence reagents with indisputable facts from homology searches. Figure 3 illustrates the kind of errors that can be detected. From a given publication, Seek & Blastn automatically extracts gene identifiers and nucleotide sequences using named entity recognition techniques. The sentence containing each sequence is automatically analyzed to assign a claimed status (targeting or non-targeting) that is compared with the most likely status according to a standard homology search.

Preliminary use of Seek & Blastn suggests that the incorrect use of nucleotide sequence reagents may be frequently undetected and represents an underestimated source of error in life science publications [29]. Following this, J. Byrne

[8] `https://www.springer.com/gp/about-springer/media/press-releases/corporate/scidetect/54166`

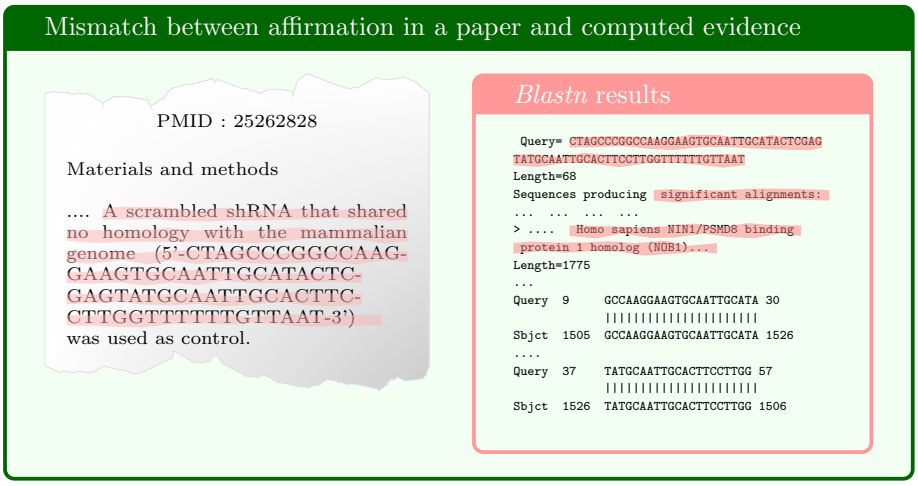


Fig. 3: Kind of errors that can be detected using the Seek & Blastn tool [27]. A nucleotide sequence is said to have no homologies with the mammalian genome but has significant homologies with the human genome (using Blastn [28]).

was named by nature as one of the 2017 Nature'10: 10 peoples who mattered in science that year[9].

4 Conclusion

Through several emblematic case studies, we showed evidences of attempts to game indicators. The presented automatic ways to detect these attempts may be seen has using on the one hand bibliometrics techniques (citation analysis) and/or information retrieval techniques on the other hand. This may not be too surprising as one of the seminal goal of scientometrics is to be able to detect and retrieve the most pertinent documents for a given set of users. The field is based on citations analysis stating, as predicate, that citations are the means by which the most pertinent documents can be identified. Often, for information retrieval the main material is the content of documents and it is assumed that this content should be used to identified relevant documents. Having a similar goal, it can be thus expected a mutual enrichment of these two families of techniques.

As a matter of fact, such approaches are very efficient in identifying generated papers, duplicated publications, plagiarism and other kind of misconduct. But significant progress is still to be made to provide valuable support to allow peers to identify and flag scientific errors in both published and forthcoming scientific literature. This could be done by means of joint analysis of citation and text.

[9] https://www.nature.com/immersive/d41586-017-07763-y/index.html#jennifer-byrne

The developed technics would also be helpful to identifying literature that brings new knowledge and expose breakthrough technologies.

But the use of such tools is a 'quick and dirty' response to misconduct problems. The situation is like if a kind of spamming war started at the heart of science. The phenomenon is taking place precisely at the very heart of science, because knowledge diffusion is at the heart of science too. It is a spamming war, because exerting high pressure on scientists mechanically leads to too prolific and less meaningful publications even if they are not non-sense.

One can invoke the Goodhart's law or state that the act of measuring a system results in that very system being disturbed. This adage is true in physics, but also in computer science and, perhaps in scientometrics and bibliometrics: by aiming at measuring science, these approaches are perturbing scientific processes, particularly when used for management purpose. The measurement of these perturbations is also a future research track that needs deeper investigation and may be within reach of bibliometrics and information retrieval technics.

Acknowledgement. The author would like to thank G. Cabanac for useful comments, supports and intellectual stimulation.

References

1. Bhattacharjee, Y.: Saudi universities offer cash in exchange for academic prestige. Science **334**(6061) (2011) 1344–1345
2. Bornmann, L., Bauer, J.: Which of the world's institutions employ the most highly cited researchers? an analysis of the data from highlycited.com. Journal of the Association for Information Science and Technology **66**(10) (2015) 2146–2148
3. Ferguson, C., Marcus, A., Oransky, I.: Publishing: The peer-review scam. Nature **515**(7528) (2014) 480–482
4. Labbé, C.: Ike antkare, one of the great stars in the scientific firmament. International Society for Scientometrics and Informetrics Newsletter **6**(2) (June 2010) 48–52
5. Antkare, I.: Architecting E-Business Using Psychoacoustic Modalities. PhD thesis, United Saints of Earth (2009)
6. Lopez-Cozar, E.D., Robinson-Garcia, N., Torres-Salinas, D.: The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. Journal of the Association for Information Science and Technology **65**(3) (2013) 446–454
7. Beel, J., Gipp, B.: Academic search engine spam and Google Scholar's resilience against it. Journal of Electronic Publishing **13**(3) (2010)
8. Beel, J., Gipp, B., Wilde, E.: Academic search engine optimization (ASEO). Journal of scholarly publishing **41**(2) (2010) 176–190
9. Labbé, C., Labbé, D.: Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? Scientometrics **94**(1) (2013) 379–396
10. Noorden, R.V.: Publishers withdraw more than 120 gibberish papers. Nature (February 2014)
11. Analyzing e-commerce process. Volume 2014 International Conference on Advances in Communication and Computing Technologies (ICACACT)., IEEE (August 2014)

12. Hvistendahl, M.: China's publication bazaar. Science **342**(6162) (2013) 1035–1039
13. Byrne, J.A., Labbé, C.: Striking similarities between publications from China describing single gene knockdown experiments in human cancer cell lines. Scientometrics **110**(3) (Mar 2017) 1471–1493
14. Huang, W., Chen, D., Ning, L., Wang, L.: Retracted: siRNA mediated silencing of NIN1/RPN12 binding protein 1 homolog inhibits proliferation and growth of breast cancer cells. Asian Pac J Cancer **18**(10) (Oct 27 2017) 2891–2891
15. Xiong, J., Huang, T.: An effective method to identify machine automatically generated paper. In: Pacific-Asia Conference on Knowledge Engineering and Software Engineering, 2009. KESE '09. (2009) 101–102
16. Lavoie, A., Krishnamoorthy, M.: Algorithmic Detection of Computer Generated Text. ArXiv e-prints (August 2010)
17. Dalkilic, M.M., Clark, W.T., Costello, J.C., Radivojac, P.: Using compression to identify classes of inauthentic texts. In: Proceedings of the 2006 SIAM Conference on Data Mining. (2006)
18. Amancio, D.R.: Comparing the topological properties of real and artificially generated scientific manuscripts. Scientometrics **105**(3) (December 2015) 1763–1779
19. Williams, K., Giles, C.L.: On the use of similarity search to detect fake scientific papers. In: Similarity Search and Applications - 8th International Conference, SISAP. (2015) 332–338
20. Ginsparg, P.: Automated screening: ArXiv screens spot fake papers. Nature **508**(7494) (04 2014) 44–44
21. Nguyen, M., Labbé, C.: Engineering a tool to detect automatically generated papers. In: Proceedings of the Third Workshop on Bibliometric-enhanced Information Retrieval co-located with the 38th European Conference on Information Retrieval (ECIR 2016), Padova, Italy, March 20, 2016. (2016) 54–62
22. Nguyen, M., Labbé, C.: Detecting automatically generated sentences with grammatical structure similarity. In: Proceedings of the Fifth Workshop on Bibliometric-enhanced Information Retrieval (BIR) co-located with the 39th European Conference on Information Retrieval (ECIR 2017), Aberdeen, UK, April 9th, 2017. (2017) 73–84
23. Bartneck, C., Servaas, K.: Detecting h-index manipulation through self-citation analysis. Scientometrics **87**(1) (2011) 85–98
24. Herteliu, C., Ausloos, M., Ileanu, B.V., Rotundo, G., Andrei, T.: Quantitative and qualitative analysis of editor behavior through potentially coercive citations. Publications **5**(2) (2017)
25. Fister jr, I., Fister, I., Perc, M.: Toward the discovery of citation cartels in citation networks. **4** (12 2016)
26. Nuijten, M.B., Hartgerink, C.H.J., van Assen, M.A.L.M., Epskamp, S., Wicherts, J.M.: The prevalence of statistical reporting errors in psychology (1985–2013). Behavior Research Methods **48**(4) (Dec 2016) 1205–1226
27. Byrne, J.A., Labbé, C.: Fact checking nucleotide sequences in life science publications: The seek & blastn tool. In: International Congress on Peer Review and Scientific Publication, Enhancing the quality and credibility of science, Chicago (September 2017)
28. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology **215**(3) (1990) 403 – 410
29. Phillips, N.: Online software spots genetic errors in cancer papers. Nature **551**(7681) (2017) 422–423