Fact-Checking Software Detects Genetic Errors in Cancer Research Publications

Author
Enago Academy

Post Url
https://www.enago.com/academy/fact-checking-software-detects-genetic-errors-cancer-research-publications/

In early October, two scientists shared a software program that detects incorrect gene sequences in already published research experiments. Using the program, the duo identified experimental flaws in more than 60 papers within cancer research alone. Scientists Jennifer Byrne and Cyril Labbé combined their expertise in cancer-research and computer-science, to introduce the software "Seek & Blastn". The program is presently at its trial phase, available online to fellow researchers for testing and improving it. The next step is to commercialize and propose the software to publishers and journal editors.

Program Objectives

Since 2015, Byrne has worked to identify errors in human-cancer papers, notably detecting errors in gene function in five papers. On scrutiny, nucleotide sequences reported in the papers were inaccurate, resulting in the retraction of two papers. Two more papers faced retraction by the 21st November. The erroneous trend appeared on

25 other papers, prompting Byrne and Labbé to develop and implement Seek & Blastn tool.

In brief, nucleotide sequences extracted from any given paper uploaded to the software will undergo fact-based verification. Technically, the program will crosscheck extracted nucleotide sequences against a public database called Nucleotide Basic Local Alignment Search tool (BLAST). If a sequence described to target a human gene is a mismatch to the Blastn database, the error is flagged. Conversely, the semi-automated program can also detect sequences described as non-targeting, with a Blastn database match. Although limited to human sequences alone at present, the pair hopes to include sequence verification for other species as well.

Program Design

In September, the researchers presented early results of their program to the International Congress on Peer Review and Scientific Publication. The conference proceedings outlined how the program could enhance scientific quality and credibility. By design, the “seek & blastn” semiautomated tool extracts gene identifiers and nucleotide sequences (15-90 bases) using entity recognition techniques. Using finite-machines, the sentence containing each sequence is analysed automatically, to assign a claimed status compared with the blastn analysis. Google Scholar enables further assessment with any status claimed within literature. To begin with, the study identified a collection of highly similar cancer research publications (CorpusP). The study further included an additional set of 154 unknown studies (CorpusU).

Preliminary Outcomes

According to the study, 48 of 48 (100%) CorpusP and 111 of 154 (73%) CorpusU publications had nucleotide sequences extracted using Seek and Blastn. The tool flagged incorrect nucleotide sequences in 38 of 48 (79%) CorpusP studies. Furthermore, the tool indicated that predictions for nontargeting sequences were incorrect compared to targeting sequences. Of the CorpusU studies, the tool flagged 30 of 154 (19%) nucleotide sequences as incorrect. Although the percentage of anomalies was small, it demonstrated substantial error for a fact-checking program nevertheless. As a result, papers put through the software also require additional manual checking at present. The program also risks complications in targeting sequence analysis, when gene identifier variations are at play. Presently, Seek & Blastn is therefore only a pilot program that seeks peer-review through follow-up analyses in the Life Sciences.

Overview of Scientific Credibility

Despite its preliminary status, the software has highlighted some core problems within the existing publications. For instance, verified sequence mismatches in publications at present can thus invalidate the results and conclusions of the paper overall. Incorrect sequence identification may indicate that results in the paper are not a true reflection of the original experiments conducted. The pair of scientists also published a study earlier

in 2016, reporting 48 problematic papers. Evidently, the papers similarly described a single gene knockdown protocol in cancer cell lines, though the authors did not perform it. In total, these two researchers have identified incorrect sequences in over 90 published papers. Such error has serious implications to the validity of ongoing research and future reproducibility of methods.

Statistician David Allison believes on using 'Seek & Blastn' to promote good scientific practice rather than to "catch people out". The expectation is that similar tools will enable error rate quantification, to regulate the reproducibility crisis. Moreover, Labbé previously identified 120 "gibberish papers" for their eventual withdrawal. Ultimately, the intention is for publishers to use the Seek & Blastn tool as part of the article-screening process. Based on the cost of the software, its accuracy and ease of use, it would enhance existing standards of academic publishing. Given the early risk for software-based error however, academic editors could explore the trial version of the software to improve it for subsequent commercialization.

Cite this article

Enago Academy, Fact-Checking Software Detects Genetic Errors in Cancer Research Publications. Enago Academy. 2017/12/06. https://www.enago.com/academy/fact-checking-software-detects-genetic-errors-cancer-research-publications/