Retrieving the evidence of a free text annotation in a scientific article: a data free approach

Julien Gobeill[1,2], Emilie Pasche[1] and Patrick Ruch[1,2]

[1] SIB Text Mining group, Swiss Institute of Bioinformatics, Geneva, Switzerland
[2] HES-SO / HEG Geneva, Information Sciences, Geneva, Switzerland
julien.gobeill@hesge.ch

Abstract. The exponential growth of research publications provides challenges for curators and researchers in finding and assimilating scientific facts described in the literature. Therefore, services that sup-port the browsing of articles and the identification of key concepts with minimal effort would be beneficial for the scientific community. Reference databases store such high value scientific facts and key concepts, in the form of annotations. Annotations are statements assigned by curators from an evidence in a publication. Yet, if annotated statements are linked with the publication's references (e.g. PubMed identifiers), the evidences are rarely stored during the curation process. In this paper, we investigate the automatic relocalization of biological evidences, the Gene References Into Function (GeneRIFs), in scientific articles. GeneRIFs are free text statements extracted from an article, and potentially reformulated by a curator. De facto, only 33% of geneRIFs are copy-paste that can be retrieved by the reader with the search tool of his reader. For automatically retrieving the other evidences, we use an approximate string matching algorithm, based on a finite state automaton and a derivative Levenshtein distance. For evaluation, two hundred candidate sentences were evaluated by human experts. We present and compare results for the relocalization in both abstracts and fulltexts. With the optimal setting, 76% of the evidences are retrieved with a precision of 97%. This data free approach does not require any training data nor a priori lexical knowledge. Yet it remarkable how it handles with complex language modifications such as reformulations, acronyms expansion, or anaphora. In the whole MEDLINE, 350,000 geneRIFs were retrieved in abstracts, and 15,000 in fulltexts ; they are currently available for highlighting in the Europe PMC literature browser.

Keywords: Natural Language Processing, Information Retrieval, Approximate string matching.

1 Introduction

Structured databases have become important resources for integrating and accessing scientific facts [1]. Yet, the normalized and integrated content still lags behind the current knowledge contained in the literature [2, 3]. Entities' properties of, such as gene functions, are usually characterized in experiments conducted by re-search teams, then reported in natural language published in scientific articles. These properties need to be

located, extracted and then integrated in normalized annotations by curators of reference gene databases, in order to be exploited by other researchers or databases. In biology, reference databases are the Gene Database maintained by the National Center for Biotechnology Information (NCBI) [4], or the UniProt database maintained by the Swiss Institute of Bioinformatics [5]. Manual curation of these scientific articles is labor intensive, but produces consistent and high-quality annotations for populating reference biological databases. In 2007, it was estimated that, despite the fact that the mouse genome is now fully sequenced, its functional annotation will not be complete in databases before 2047 [6].

Statements about gene functions, known as Gene References into Function (GeneRIFs), are collected by the NCBI. They are short statements extracted by life science experts from scientific articles. GeneRIFs are intended to facilitate access to publications documenting experiments on gene functions. Yet, geneRIFs are simply linked to an article, but not localized in the fulltext. This loss of information during the curation process is harmful, as curators want to learn about new functions as quickly as they can, preferably without having to scan all the paper for retrieving the evidence. GeneRIFs relocalization can thus be performed by automated approaches in order to help curators to keep up with the growing flow of publications.

Nowadays, machine learning and expert systems are popular for automatic tasks in biomedical articles [7, 8]. Yet, both approaches need a critical amount of a priori knowledge, respectively learning data and hand-written language rules, before being able to manage edit modifications or reformulations. Such a priori knowledge can be costly to gather, when it is sometimes simply not available. In contrast, approximate string matching provides a data-free approach for estimating the similarity between an evidence and a passage.

In this paper, we investigate the abilities of an approximate string matching algorithm for retrieving evidences (geneRIFs) in scientific publications. This paper is organized as follows. Section 2 reports related works on approximate string matching in biomedical literature, and on geneRIFs retrieval. Section 3 describes the data, the methods, and the benchmarks used for evaluation. Lastly, section 4 reports evidence retrieval results on abstracts, then on fulltexts.

2 Related work

In biology, approximate string matching is a popular approach for gene name recognition [9]. For this specific task, survey studies compare and optimize different algorithms in terms of performance and computation time [10]. Approximate string matching was also investigated for more general named entity recognition [11], alignment of DNA sequences [12], optical character recognition [13], or approximate text search in literature [14].

Dealing with geneRIFs, the Genomics Track in the Text Retrieval Conferences (TREC) addressed in 2003 the issue of extracting geneRIFs statements from a scientific publication [15]. Participating teams were asked to maximize the lexical overlap between the geneRIF and a passage, measured by the Dice coefficient. The track overview

reported two best approaches. The first team [7] investigated sentences normalization thanks to stemming, gene names dictionaries and thesaurus, then trained a Bayesian classifier. The second team [16] also trained a Bayesian classifier, but input sentences were only abstracts title and last sentences, and features were only normalized gene names and verbs. No groups obtained much improvement compared to the baseline, which consisted in choosing titles. Yet, the lexical overlap was pointed out as a weak measure of equivalence, as geneRIFs can be para-phrases of articles sentences.

In [17], generated evidences were compared to authentic geneRIFs. For text representation, authors produced several features based on sentence position, sentence discourse, gene normalization or ontology terms mapping. Furthermore, 3,000 sentences were annotated by experts for building training and test sets. Despite these numerous efforts, authors concluded that their machine learning approach did not produce results comparable to sentence position.

3 Methods

In this section, we first focus on data: the geneRIFs dataset, and the corresponding articles (abstracts accessed via MEDLINE, and fulltexts accessed via PubMed Central). We then deal with the methods: we introduce the Levenshtein distance, and the derivative distance used on this work. Finally, we detail the evaluation: benchmarks, judgements and metrics

3.1 GeneRIFs

GeneRIFs are freely available in the NCBI server (ftp://ftp.ncbi.nih.gov/gene/GeneRIF/). The whole dataset was acquired on August 2018, and contained 1.2 million of geneRIFs.
According the NCBI website, geneRIFs allow any scientist to add the functional annotation of a gene contained in the Gene database. The geneRIFs must be linked to an existing gene entry. Three information are mandatory for completing a submission:

1. A concise phrase describing a function (less than 425 characters in length).
2. A published paper describing that function, implemented by supplying the PubMed identifier (PMID) of a citation in MEDLINE.
3. The curator’s e-mail address.

The concise phrase is free text : the geneRIF curator is free to copy-paste, edit or reformulate the evidence described in the paper. Moreover, a geneRIF is linked to a publication, not to the specific passage that describes the function. Yet, it is stated that the title is not accepted.

Once submitted, the text of the GeneRIF is reviewed for inappropriate content and typographical errors, but not otherwise edited. Finally, it is stated that most of the GeneRIFs are provided by the staff of the National Library of Medicine's Index Section, who have advanced degrees in the life sciences.

Here is one example of geneRIF in json format: {“geneID”: 2827861, “PMID”: 15664975, “text: “Strains with mutations in either the prcA-prtP or the msp region showed altered expression of the other locus. (msp = major outer sheath protein)”}. In this example, the geneRIF annotator has copy pasted a sentence from the abstract, and has developed the acronym “msp” (which is not developed in the abstract).

3.2 Publications

All geneRIFs are provided with a PubMed identifier (PMID), which is linked with a specific publication contained in the MEDLINE database. Thus, for all papers referred in geneRIFs, titles and abstracts are freely accessible via MEDLINE. The open access fulltexts can be freely accessed via the PubMed Central database.

Abstracts from MEDLINE. MEDLINE is a bibliographic database of life sciences. In January 2018, it contained 27 million records of scientific papers from more than 5,500 selected journals in the domains of biology and health. MEDLINE is maintained by the United States National Library of Medicine (NLM), and is accessible via the NLM search engine (PubMed), or freely downloadable via services and FTP.

A MEDLINE record contains the title and the abstract, and also metadata such as authors’ information, journal’s information, the publication date, or keywords (Medical Subject Headings) added by the NLM’s indexers. All MEDLINE records are uniquely identified by a PMID.

Fulltexts from PubMed Central. PubMed Central is a free fulltext database of publicly accessible literature published in life sciences. In January 2018, it contained 4.6 million records. Yet, only a subset of 1.4 million of publications is Open Access and linked with a PMID. 444,000 geneRIFs are linked with a paper contained in PubMed Central. Out of these 444,000, only 153,000 (35%) are linked with a paper contained in the Open Access subset.

The BioMed platform. For accessing papers, we used BioMed, a local resource for literature access and enrichment. BioMed provides access to Open Access abstracts and fulltexts, thanks to a synchronized mirror of MEDLINE and PubMed Central. It also provides some text services such as information extraction [18], question answering [19, 20], or sentence splitting. BioMed is currently used in the workflow of protein curators at the Swiss Institute of Bioinformatics [21].

3.3 Algorithm

The Levenshtein distance. The Levenshtein distance [22] quantifies the similarity between two strings ; for this purpose, it computes the number of single-characters operations required to transform one string into the other. The available operations are:

- d: deletion of a character
- s : substitution of a character with another
- i: insertion of a character

For instance, three operations are requested for transforming the string “kitten” into the string “sitting”: (1) substitution of “k” with “s”, (2) substitution of “e” with “i”, (3) insertion of “g”. The number of operations is seen as the distance: the less operations are requested, the most similar strings are. An exact match results in a distance of zero. Finally, each operation can be weighted in order to tune the algorithm for a specific task.

The Evidence Retrieval distance. Our approach for retrieving a geneRIF in a given paper is sentence-centric. First, the abstract (or fulltext when it is available) is split into sentences, thanks to local services provided by the BioMed platform. Then, all sentences are compared with the geneRIF using a derivative Levenshtein distance. There are different available implementations of the Levenshtein algorithm, including recursive, or iterative with matrix [23]. For this study, we have used a library from the Comprehensive Perl Archive Network (Text::Fuzzy).

Preliminary analysis revealed some limitations in the default Levenshtein distance for evidence retrieval. First, the deletion of characters is actually not an issue: a lot of geneRIFs curators choose to cut several words from a sentence when producing the statement (such as “We show that” or “These results indicate that”). We have thus decided to ignore deletion operations. Second, the Levenshtein distance has no normalization according to the string length: a number of ten operations is obviously better for a thirty words evidence than for a five words one. We have thus introduced a normalization with the total number of characters in the geneRIF. We finally obtained a derivative Levenshtein distance, called Evidence Retrieval (ER) distance, as given below:

$$ER\ Distance(str_1, str_2) = \frac{S + I}{length(str_2)}$$

In this formula, str1 is a candidate sentence, str2 is the evidence to retrieve, S and I are the total number of substitutions and insertions required to transform str1 into str2 (deletions are not counted), and length(str2) is the number of characters in str2. The ER distance between a geneRIF and a candidate sentence can be seen as the percentage of characters in the geneRIF that required to be introduced or substituted in the sentence. An ER distance of zero means that the geneRIF is contained in the sentence, but that the sentence may contain some supplementary characters that were not chosen by the curator.

3.4 Evaluation

Test sets. For evaluation purposes, we designed two test datasets of 100 geneRIFs. The first test set was sampled in the set of the 444,000 geneRIFs linked to publications available in MEDLINE. The task evaluated with this first test set was evidence retrieval

in abstracts. The second test set was sampled in the set of 153,000 geneRIFs linked to publications contained in the PMC Open-Access subset. The task evaluated with this second test set was evidence retrieval in fulltexts.

For all geneRIFs in the benchmarks, ER distances were computed between the geneRIF and all publication's sentences. Then, only the sentence with the smallest ER distance was selected: we call it the candidate sentence. Results were analyzed according intervals and thresholds of ER distance values.

Equivalence judgements. Two experts evaluated the equivalence between the geneRIFs and the candidate sentences. Experts were bioinformaticians, one having an advanced degree in information sciences, the other one in biology. They were asked to judge the equivalence between a geneRIF and a candidate sentence, according to three possible values:

- 100%: the candidate sentence contains all the annotatable in-formation contained in the geneRIF. The expert thinks that a curator could make an annotation only with this sentence.
- 50%: the candidate sentence contains some annotatable in-formation contained in the geneRIF. The expert thinks that this sentence could help a curator, but also that other useful information are contained in other sentences.
- 0%: the candidate sentence contains no annotatable information contained in the geneRIF. The expert thinks that this sentence would be of no help for a curator.

The pairs (geneRIF, candidate sentence) were presented to the experts in a random order. ER distances were hidden from experts. For each candidate sentence, the final equivalence is the average judgement of both experts. For instance, if one expert gave a 100% equivalence and the other one 50%, the final retained equivalence for this candidate sentence is 75%. Moreover, for each candidate sentence, an Inter-Annotator (IA) Agreement was computed. The IA Agreement is 100% minus the absolute difference between both judgements. For instance, if one expert gave a 100% equivalence and the other one 50%, the final IA Agreement for this candidate sentence is 50%.

4 Results & Discussion

First of all, we present some preliminary results and statistics on geneRIFs and papers' sentences. Next, we focus on evidence retrieval in abstracts, then on evidence retrieval in fulltexts. For each benchmark, we present some interesting examples of candidate sentences and edit modifications handled by our approach.

4.1 Preliminary results

GeneRIFs' length. We used regular expressions in order to compute and compare the lengths of geneRIFs, and the lengths of sentences in an abstract or a fulltext. Lengths

were computed with samples of 20,000 geneRIFs, 4,000 abstracts and 400 fulltexts. Mean and quartiles are showed in Table 1.

Table 1. Length of geneRIFs, and sentences in abstracts and fulltexts.

	Mean	1st q.	Median	3rd q.
GeneRIFs	21	14	20	27
Abstracts sentences	24	16	22	29
Fulltexts sentences	20	9	18	33

In terms of length, GeneRIFs are quite similar to sentences from abstracts. For fulltexts, mean and median are similar to geneRIFs, but the distribution is sparser: there seems to be more short and long sentences in fulltext. Finally, we observe a mean of ten sentences per abstract, versus five hundred per fulltext. This strengthens our sentence-centric approach. The sparser distribution in fulltexts could be explained by the fact that abstracts sentences aim at summarizing the fulltext content. Fulltext sentences are thus more likely to be short or long (such as long explanations). In this perspective, geneRIFs seem to be more factual evidences than detailed explanations.

GeneRIFs' words presence in abstracts. We then studied what proportions of geneRIFs words can be found in abstracts. In Figure 1, 440 sampled geneRIFs are plotted according to their proportion of words present in the corresponding abstract.

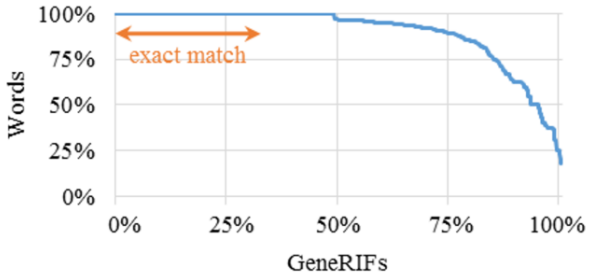

Fig. 1. GeneRIFs words present in abstracts.

50% of geneRIFs have all their words contained in the abstract, including 33% that are found in exact match. The next 25% have between 90% and 100% of their words in the abstract, while the proportion quickly drops for the last 25%. This leads to several assumptions. First, a huge proportion of geneRIFs seems to be extracted from the abstracts. Second, simple exact match already retrieves one third of the evidences, which seems to be a fair baseline. The second third has above 95% of words present in the abstract; these evidences could be retrieved by approximate string matching. Finally,

the last third could be less reachable by our approach (at least for abstracts), as fewer words are present.

4.2 Evidence retrieval in abstracts

We now present the results of evidence retrieval in abstracts. For one hundred geneRIFs, two experts were asked to judge the equivalence of the best candidate sentence extracted from the corresponding abstract.

Equivalence according to distance intervals. Equivalences and IA Agreements are given in Fig. 2, according to ER distance intervals.

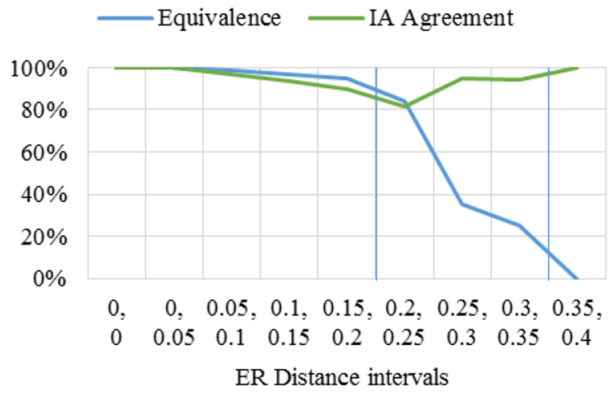

Fig. 2. Equivalence according to distance intervals.

It is remarkable how the equivalence curve quickly drops. We can split the curve into three parts (boundaries are vertical blue lines):

- for values between 0 and 0.2, equivalence is above 95% (even 100% for [0, 0.05] values)
- then, between 0.2 and 0.35, equivalence drops from 84% to 35% and 25%
- finally, for distances bigger than 0.35, equivalence falls to 0%.

It is also remarkable how the IA Agreement curve behaves according to these three parts. For the first and last ones, the agreement is very high (above 90%, and even 100% for 100% or 0% equivalence), while the central part seems to be more uncertain for judges.

Thus, the ER distance shows high abilities to produce two distinct sets of true positives (100% equivalence) and true negatives (0% equivalence), with 100% agreement in both cases. Between both is a set of more uncertainty for middle ER distances, with smaller equivalence values and less agreement between judges.

Equivalence according to distance thresholds. We know consider equivalence according to proportion of retrieved geneRIFs. We did not focus anymore on ER distance intervals but on ER distance thresholds, in order to know what proportion of geneRIFs are retrieved in abstracts with a given threshold. Results are given in Fig. 3.

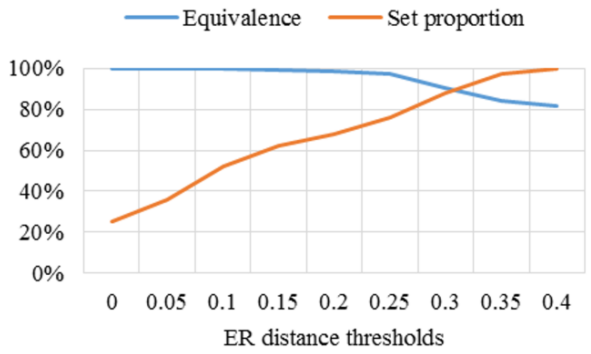

Fig. 3. Equivalence and set proportion for abstracts.

Considering the previous first part, 68% of geneRIFs are retrieved with an equivalence mean of 99% with an ER distance threshold of 0.2. With a threshold of 0.25, 76% of geneRIFs are retrieved with an equivalence mean of 97%. Considering the drop in Figure 2, these thresholds are to be considered for delivering an optimal output.

Interesting examples. We now present some remarkable retrieved evidences.

Words addition or deletion. Beyond copy-pastes, the geneRIFs curators are likely to delete some unnecessary words, or to add words in order to bring some precision that is not explicit in the sentence.

GeneRIF: “PfSir2a fine-tunes ribosomal RNA gene transcription.” Abstract sentence (ER distance 0): “Here we investigate the nucleolar function of PfSir2a and demonstrate that PfSir2a fine-tunes ribosomal RNA gene transcription.” In this case, the geneRIF curator chose to cut the evidence introduction.

GeneRIF: “Transgenic flies, expressing the human ERRa-G allele, constitutively over-express Cyp12d1, Cyp6g2 and Cyp9c1.” Abstract sentence (ER distance 0.05): “Transgenic flies, expressing the ERRa-G allele, constitutively over-expressed Cyp12d1, Cyp6g2 and Cyp9c1.” In this case, the geneRIF curator simply adds “human” in order to add context and to specify the related organism.

Acronyms expansion. Authors often use acronyms, which are developed only the first time they mention them in the article (e.g. “EC” for “Endothelial cells”). Thus, geneRIFs curators are likely to develop the acronyms used in the extracted evidence. Moreover, Greek letters are often used in gene names, while geneRIF curator can prefer

full letter names (e.g. “alpha” instead of “α”). These kinds of modifications can be handled by pattern substitutions and numerous rules, especially for acronym expansions [24]. In contrast, approximate string matching can handle with these modifications without linguistic knowledge.

GeneRIF: “interaction of alpha6beta1 in embryonic stem cells (ECCs) with laminin-1 activates alpha6beta1/CD151 signaling which programs ESCs toward the endothelial cells lineage fate” Abstract sentence (ER distance 0.28): “Thus, interaction of α6β1 in ESCs with LN1 activates α6β1/CD151 signaling which programs ESCs toward the EC lineage fate.” In this case, the geneRIF curator developed two Greek letters, and three acronyms. Even if the ER distance is quite high, it is remarkable how approximate string matching dealt with so many substitutions and selected the good sentence.

Anaphora resolution. As evidences can deal with just mentioned patterns in the argumentative flow, authors are likely to use pronouns (e.g. “it”) or anaphora instead of writing repetitions.

GeneRIF: “Under hypoxia reoxygenation or ischemia and reperfusion, StAR and CYP11A1 protein and gene expression was reduced without apparent relation to TSPO changes”. Abstract sentence (ER distance 0.20): “Under the same conditions, StAR and CYP11A1 protein and gene expression was reduced without apparent relation to TSPO changes”. In this case, the author did not repeat the experiment conditions in his sentence; highlighting the evidence in the paper allows a curator to quickly check in the neighborhood of the sentence the sentence if the reported conditions are the same.

4.3 Evidence retrieval in fulltexts

We now present the results of evidence retrieval in fulltexts.

Equivalence according to distance intervals. Equivalences and IA Agreements are given in Fig. 4, according to ER distance intervals.

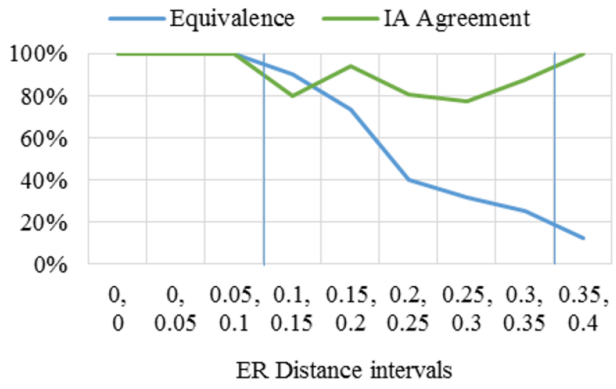

Fig. 4. Equivalence according to distance intervals.

Equivalence curves share the same shape for fulltexts than previously for abstracts:

- for values between 0 and 0.1, equivalence is 100%
- then, between 0.1 and 0.35, equivalence slowly drops from 90% to 25%
- finally, for distances bigger than 0.35, equivalence falls to 13%.

As for abstracts, the IA Agreement is 100% when equivalence is maximum or minimum. Yet, for the central part, we observe more uncertainty for judges in order to determine if the sentence are equivalent or not.

Equivalence according to distance thresholds. Results linking ER distance thresholds and proportions of retrieved geneRIFs in fulltexts are given in Fig. 5.

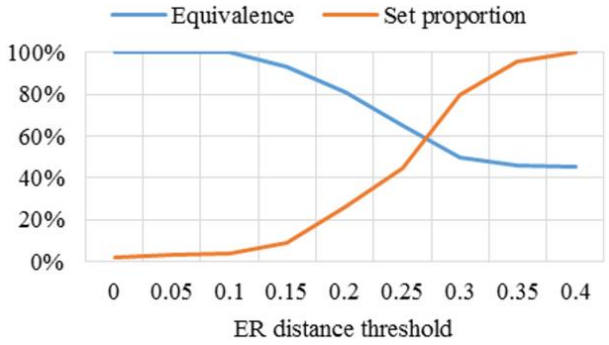

Fig. 5. Equivalence and set proportion for abstracts.

Considering the previous first part, only 4% of geneRIFs are retrieved in fulltext with an equivalence of 100% with an ER distance threshold of 0.1. Going up to a threshold of 0.15, 9% of geneRIFs are retrieved with an equivalence mean of 90%. Considering the drop observed in Figure 4, these thresholds seem to be considered for delivering an optimal output. Bigger values for thresholds could be risky in a non-reviewed production.

It is difficult to evaluate the complementarity of evidence retrieval in fulltexts and abstracts. Further analyses reveal that for the five true positives in fulltexts with an ER distance inferior to 0.1, four were also true positives in abstracts (with an average ER distance of 0.03), while only one was a true negative in the abstract (with an ER distance of 0.32). GeneRIFs curators tend to preferably choose their evidence in abstracts. For geneRIFs that were missed in abstracts, a small portion could be retrieved in fulltext.

Another aspect to consider is that the abstract is a summary of fulltext. Thus, when the geneRIFs are created from an abstract sentence, there is probably a passage in the fulltext where the evidence comes from, probably differently formulated, and more detailed. This could explain why ER distances are larger in fulltexts than in abstracts. In

other words, the approximate string matching would retrieve fulltext sentences that were summarized in the abstract and used as geneRIFs.

Yet, for approximately one quarter of geneRIFs, our approach is not able to retrieve the evidence neither in the abstract nor in the fulltext. In the below section, we present some of these missed geneRIFs.

Interesting examples. We now present some remarkable retrieved evidences.

Reformulated evidences. GeneRIF: “Three FNR proteins (ANR, PP_3233, and PP_3287) respond to O2 differently.” Fulltext sentence (ER distance 0.15): “Thus, the response of ANR was similar to that reported previously for E. coli FNR, further confirming the similarities between these two proteins, but PP_3233 and PP_3287 were less responsive with both O2 and nitric oxide compared with ANR (25).” In this case, the sentence contains the same information than the geneRIF, and was selected by our algorithm even if the interaction was interpreted and reformulated into “respond to O2 differently”.

GeneRIF: "These data support the role of proconvertase PCSK5 in the processing of ovarian inhibin subunits during folliculogenesis." Fulltext sentence (ER distance 0.18): "We demonstrate that the spatial and temporal expression of the proconvertase PCSK5 overlaps with the expression and processing of mature inhibin subunits in the ovary during follicle expansion." In this case, a good sentence was retrieved, while “we demonstrate that the […] expression of the proconvertase PCSK5 overlaps…” was reformulated into “These data support the role of proconvertase PCDK5 in…“.

4.4 Missed geneRIFs

Here are some examples of missed geneRIFs.

GeneRIFs: “Observational study and genome-wide association study of gene-disease association.” “Clinical trial of gene-disease association and gene-environment interaction” “Observational study of gene-disease association, gene-environment interaction, and pharmacogenomic / toxicogenomic.”

These three examples taken in the abstracts benchmark are more descriptions of the studies than facts REWRITE. Thus, it is not surprising that no good sentence was retrieved in the articles.

In other geneRIFs that were not successfully retrieved, we observed several long geneRIFs with multiple sentences, probably gathering facts from different parts of the article. At last, some geneRIFs are statements inferred by the curator, thus are beyond the scope of statistical or linguistic systems.

5 Conclusion

We investigated approximate string matching for the task of retrieving evidences expressed in geneRIFs in their corresponding papers. We defined an Evidence Retrieval distance derived from the Levenshtein distance. When selecting the abstract sentence

having the smallest distance value, this approach can retrieve 76% of evidences with a very high precision (97%). While 33% of geneRIFs are just copy-pastes, it is remarkable how this approach can handle with complex editing modifications, including reformulations, acronym expansions, or anaphora resolutions. Beyond evidence retrieving, approximate string matching could be used in fulltext for retrieving sentences that contain the information summarized in an abstract.

It is worth reminding that, in contrast with machine learning or rule-based approaches, which need a substantial amount of training data or a priori knowledge, such classic approaches are on the shelf, and are still very effective for concrete text mining applications.

In 2017, within the Elixir Excelerate project, which aims at ensuring the delivery of world-leading life-science data services, the whole geneRIFs dataset was treated. 337,000 were retrieved in abstracts, and 13,770 in fulltexts. These retrieved geneRIFs were then used in order to fill the Europe PMC database [25,1], maintained by the European Bioinformatics Institute (EBI). These geneRIFs are now available for highlighting when a user reads an article. We hope that this will help the curators to navigate in the literature in their daily workflow, thus facilitating to bridge the gap between information contained in literature and in databases.

Acknowledgments. This research was supported by the Elixir Excelerate project, funded by the European Commission within the Research Infrastructures programme of Horizon 2020, grant agreement number 676559. The authors thank their colleagues from the SIB Swiss Institute of Bioinformatics (Core-IT), in particular Daniel Texeira and Heinz Stockinger, who provided insight and expertise that greatly assisted the research. The authors also thank the European Bioinformatics Institute, in particular Johanna McEntyre and Aravind Venkatesan, for the integration of retrieved geneRIFs into EuropePMC.

References

1. Venkatesan, A., Kim, J. H., Talo, F., Ide-Smith, M., Gobeill, J., Carter, J., et. al. 2016. SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data. Wellcome Open Research, 1. DOI = http://dx.doi.org/10.12688/wellcomeopenres.10210.1
2. Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., et. al. 2008. Big data: The future of biocuration. Nature, 455,7209, 47-50. DOI = http://dx.doi.org/10.1038/455047a
3. Gobeill, J., Pasche, E., Vishnyakova, D., Ruch, P. 2013 Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases. Database (Oxford). DOI = https://doi.org/10.1093/database/bat041
4. Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., et. al. 2015. Gene: a gene-centered information resource at NCBI. Nucleic acids research. 43, D1, D36-D42. DOI = https://doi.org/10.1093/nar/gku1055
5. Bultet, L. A., Aguilar-Rodriguez, J., Ahrens, C. H., Ahrne, E. L., Ai, N., et. al. 2016. The SIB Swiss Institute of Bioinformatics" resources: focus on curated databases. Nucleic acids research, 44, D27-D37. DOI = https://dx.doi.org/10.1093/nar/gkv1310

6. Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., and Hunter, L. 2007. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics, 23, 13, i41-i48. DOI = https://dx.doi.org/10.1093%2Fbioinformatics%2Fbtm229
7. Jelier, R., Schuemie, M. J., van der Eijk, C. C., Weeber, M., Van Mulligen, E. M., Schijvenaars, B. J., et. al. 2003. Searching for geneRIFs: Concept-Based Query Expansion and Bayes Classification. In TREC Proceedings, 225-233
8. Obermeyer, Z., and Emanuel, E. J. 2016. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. The New England journal of medicine. 375, 13, 1216. DOI = https://dx.doi.org/10.1056%2FNEJMp1606181
9. Tsuruoka, Y., and Tsujii, J. I. 2004. Improving the performance of dictionary-based approaches in protein name recognition. Journal of biomedical informatics, 37, 6, 461-470
10. Papamichail, D., and Papamichail, G. 2009. Improved algorithms for approximate string matching. BMC bioinformatics, 10, 1, S10
11. Wang, W., Xiao, C., Lin, X., and Zhang, C. 2009. Efficient approximate entity extraction with edit distance constraints. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, 759-770
12. Buschmann, T., Bystrykh, L. V. 2013. Levenshtein error-correcting barcodes for multiplexed DNA sequencing. BMC bioinformatics, 14, 1, 272
13. Lasko, T. A., & Hauser, S. E. 2000. Approximate string matching algorithms for limited-vocabulary OCR output correction. In Photonics West 2001-Electronic Imaging, 232-240
14. Wang, J., Cetindil, I., Ji, S., Li, C., Xie, X., Li, G., and Feng, J. 2010. Interactive and fuzzy search: a dynamic way to explore MEDLINE. Bioinformatics. 26, 18, 2321-2327
15. Hersh, W. R., & Bhupatiraju, R. T. 2003. TREC genomics track overview. In TREC Proceedings, 14-23
16. Bhalotia, G., Nakov, P., Schwartz, A. S., and Hearst, M. A. 2003. BioText Team report for the TREC 2003 Genomics Track. In TREC Proceedings, 612-621
17. Jimeno-Yepes, A. J., Sticco, J. C., Mork, J. G., and Aronson, A. R. 2013. GeneRIF indexing: sentence selection based on machine learning. BMC bioinformatics, 14, 1, 171
18. Gobeill, J., Ruch, P., Zhou, X. 2008. Query and document expansion with medical subject headings terms at medical imageclef 2008. In Workshop of the Cross-Language Evaluation Forum for European Languages, 736-743. Springer Berlin Heidelberg
19. Gobeill, J., Gaudinat, A., Pasche, E., Vishnyakova, D., Gaudet, P., Bairoch, A., Ruch, P. 2015. Deep Question Answering for protein annotation. Database (Oxford). DOI= http://dx.doi.org/10.1093/database/bav081
20. Pasche, E., Teodoro, D., Gobeill, J., Ruch, P., & Lovis, C. 2009. QA-driven guidelines generation for bacteriotherapy. In AMIA Annual Symposium Proceedings, 509-513
21. Mottin, L., Gobeill, J., Pasche, E., Michel, P. A., Cusin, I., Gaudet, P., & Ruch, P. 2016. neXtA5: accelerating annotation of articles via automated approaches in neXtProt. Database, baw098
22. Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady 10, 8, 707-710
23. Wagner, R. A., & Fischer, M. J. 1974. The string-to-string correction problem. Journal of the ACM (JACM), 21, 1, 168-173
24. Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M., & Morrell, M. 2001. Automatic extraction of acronym-meaning pairs from MEDLINE databases. Studies in health technology and informatics, 1, 371-375
25. Europe PMC Consortium. 2014 Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic acids research. DOI = https://doi.org/10.1093/nar/gku1061