Journal of Biomedical Informatics 71 (2017) 229–240

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Automated detection of records in biological sequence databases that are inconsistent with the literature

Mohamed Reda Bouadjenek *, Karin Verspoor, Justin Zobel

Department of Computing and Information Systems, The University of Melbourne, Parkville 3053, Australia

ARTICLE INFO

Article history:
Received 15 January 2017
Revised 9 June 2017
Accepted 12 June 2017
Available online 15 June 2017

Keywords:
Data analysis
Data quality
Bioinformatics databases
Anomaly detection

ABSTRACT

We investigate and analyse the data quality of nucleotide sequence databases with the objective of automatic detection of data anomalies and suspicious records. Specifically, we demonstrate that the published literature associated with each data record can be used to automatically evaluate its quality, by cross-checking the consistency of the key content of the database record with the referenced publications. Focusing on GenBank, we describe a set of quality indicators based on the relevance paradigm of information retrieval (IR). Then, we use these quality indicators to train an anomaly detection algorithm to classify records as "*confident*" or "*suspicious*".

Our experiments on the PubMed Central collection show assessing the coherence between the literature and database records, through our algorithms, is an effective mechanism for assisting curators to perform data cleansing. Although fewer than 0.25% of the records in our data set are known to be faulty, we would expect that there are many more in GenBank that have not yet been identified. By automated comparison with literature they can be identified with a precision of up to 10% and a recall of up to 30%, while strongly outperforming several baselines. While these results leave substantial room for improvement, they reflect both the very imbalanced nature of the data, and the limited explicitly labelled data that is available. Overall, the obtained results show promise for the development of a new kind of approach to detecting low-quality and suspicious sequence records based on literature analysis and consistency. From a practical point of view, this will greatly help curators in identifying inconsistent records in large-scale sequence databases by highlighting records that are likely to be inconsistent with the literature.

1. Introduction

Bioinformatics sequence databases such as GenBank or UniProt contain large numbers of nucleic acid sequences and protein sequences. In 2017, GenBank alone contained over 228 billion nucleotide bases in more than 199 million sequences – a number that is growing at an exponential rate, doubling every 18 months.[1] In commercial organizations, the primary reason for creating and maintaining such databases is their importance in the process of drug discovery, while in research they are used to understand the biological basis of disease. Thus, a high level of data quality is crucial.

However, since these databases are fed by direct submissions from individual laboratories and by bulk submissions from large-scale sequencing centers, they suffer from a range of data quality issues [1] including errors, redundancies, ambiguities, incompleteness, and as we will show, discrepancies such as inconsistency with the literature. Most of these records are linked to research articles in which the sequence was reported, but the need to manually create the records on such a large scale means that errors creep in and, given the volume, human curation alone is not sufficient for detection of these errors.

In this work, we seek to investigate and analyse the data quality of sequence databases from the perspective of a curator, who must detect anomalous and suspicious records. In contrast to previous research, which has concerned detection of duplicate records [2–4] and erroneous annotations [5–7], we emphasize detection of low-quality records that we define as being inconsistent with the published literature. Specifically, we propose that the literature that is linked to records in their "reference" fields be automatically used as background knowledge to check their quality. We explore a combination of information retrieval (IR) and machine learning techniques to identify records that are anomalous and thus merit analysis by a curator.

* Corresponding author.
E-mail addresses: reda.bouadjenek@unimelb.edu.au (M.R. Bouadjenek), karin.verspoor@unimelb.edu.au (K. Verspoor), jzobel@unimelb.edu.au (J. Zobel).

[1] http://www.ncbi.nlm.nih.gov/GenBank/statistics/.

http://dx.doi.org/10.1016/j.jbi.2017.06.015
1532-0464/

To provide insight into the data quality of the nucleotide records cited by articles available in PubMed Central[2] (PMC) from a literature consistency point of view, we analyzed these records as illustrated in Fig. 1. This figure shows the term overlap similarity[3] between the record definition and different sections of its associated article(s) (representing the title, abstract, body, and the full text). There are three notable trends here: first, term overlap increases from title to body and full text since the size grows accordingly; second, there is a high term overlap of roughly 80% between the record description field and the literature body section; and third, when considering the overlap similarity between the description field of the records and the full text of their associated articles, for a small number of records in which the overlap similarity is below the value of the minimum whisker (0.4), there is low overlap or no overlap at all, thus statistically suggesting a data quality problem.

As an example, the record with accession number KM403369[4] doesn't share any terms with the article PMC4465667[5] that is supposed to report on that record. Compared to the median value, which is roughly 80% similarity between a record description field and the body section of the article (see Fig. 1), this association can be considered an outlier from a statistical perspective, and can be argued to be weak. While this observation is purely statistical, it may be an indicator of a low confidence in that record. Although this record is not necessary faulty, its characteristics in relation to the overall statistical distribution clearly suggest that it should be flagged as "suspicious", and should be sent to a curator for further investigation.

Usually, a suspicious record is reported manually, by a curator whose the job consists mainly to check the database records, the record's original submitter, or a third person who may use the database and notice the inconsistency of that record. To illustrate the difficulty of the task of identifying failing records, we analysed the distribution of record ages, for records which have been removed. This analysis showed that removed records have an average age of about 1 month at their removal time. This leads us to make two hypotheses: either (i) it takes about one month for a problematic record to be detected, or (ii) curators focus only on new records, while neglecting older ones. Either way, it is clear that there is a time window of only 1 month during which curators act. Hence, if a suspicious record is not identified in this time frame, it has a low probability of being spotted. These observations show the difficulty of the curator's job, and the need for the development of automatic methods to assist them.

With the aim of assisting curators, and while focusing on GenBank, we present in this paper a method for detection of suspicious records based on their associated articles and also on the collection of articles as a whole. To the best of our knowledge, this work is the first to use the literature for data quality assessment of bioinformatics sequence databases. The contributions of this paper are as follows:

- We demonstrate that the research literature can be automatically used for assessing the quality of a record.
- We propose a list of quality indicators that correlate with the quality of a record. The quality indicators are then used to train a learning anomaly detection algorithm.
- Our experiments on the full PubMed Central collection show that, although less than 0.25% of the records in our data set are faulty, by automated comparison with literature they can be identified with a precision of up to 10% and a recall of up to 30%, while greatly outperforming the best baseline.

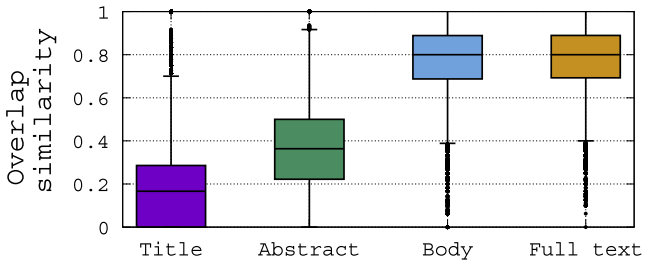

Fig. 1. Overlap similarity between a record definition field and different sections of its associated document.

2. Related work

There is a substantial body of research related to data quality in bioinformatics databases. Previous research has focused mainly on duplicate record detection and erroneous annotations, as reviewed below.

2.1. Duplicate records

Koh et al. [4] use association rule mining to check for duplicate records with per-field exact, edit distance, or BLAST sequence [8] alignment matching. Drawbacks of this method, and its poor performance, have been shown by Chen et al. [2,3]. Similarly, Apiletti et al. [9] proposed extraction of association rules among attribute values to find causality relationships among them. By analysing the support and confidence of each rule, the method can show the presence of erroneous data. Other approaches also use approximate string matching to compute metadata similarity [10–12]. However, as they focus only on metadata, the underlying interpretation is that duplicates are assumed to have high metadata similarity, or that their sequences are identical.

Other approaches consider duplicates at the sequence level; they examine sequence similarity and use a similarity threshold to identify duplicates. For example, Holm and Sander [13] identified pairs of records with over 90% mutual sequence identity. Heuristics have been used in some of these methods to skip unnecessary pairwise comparisons, thus improving the efficiency. Li and Godzik [14] proposed CD-HIT, a fast sequence clustering method that uses heuristics to estimate the anticipated sequence identity and will skip the sequence alignment if the pair is expected to have low identity. Recently, Zorita et al. [15] proposed Star Code to detect duplicate sequences, which uses the edit distance as a threshold and will skip pairs exceeding the threshold. Such methods are valuable for this task, but do not address the problem of consistency or anomaly.

2.2. Erroneous annotations

Sequence databases exist as a resource for biomedicine, but the utility of the sequence of an organism depends on the quality of its annotations [12]. The annotations indicate the locations of genes and the coding regions in a sequence, and indicate what those genes do. That is, annotations serve as a reading guide to a sequence, which makes the scientific community highly reliant on this information. Although the research and development of algorithms for identifying coding sequences (CDSs) is still an active area in bioinformatics research, genome annotation has evolved greatly during past few years [16–19]. However, the functional annotation of CDSs is particularly difficult to automate [20].

[2] http://www.ncbi.nlm.nih.gov/pmc/.

[3] We use the overlap similarity to emphasize the number of terms of a record definition that are in its associated article. Here, $Overlap(X_1, X_2) = |X_1 \cap X_2| / \min(|X_1|, |X_2|)$.

[4] http://www.ncbi.nlm.nih.gov/nuccore/KM403369.

[5] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4465667/.