Learning Biological Sequence Types Using the Literature

Mohamed Reda Bouadjenek
The University of Melbourne
School of Computing and Information Systems
Parkville, Victoria 3010, Australia
reda.bouadjenek@unimelb.edu.au

Karin Verspoor
The University of Melbourne
School of Computing and Information Systems
Parkville, Victoria 3010, Australia
karin.verspoor@unimelb.edu.au

Justin Zobel
The University of Melbourne
School of Computing and Information Systems
Parkville, Victoria 3010, Australia
jzobel@unimelb.edu.au

ABSTRACT

We explore in this paper automatic biological sequence type classification for records in biological sequence databases. The sequence type attribute provides important information about the nature of a sequence represented in a record, and is often used in search to filter out irrelevant sequences. However, the sequence type attribute is generally a non-mandatory free-text field, and thus it is subject to many errors including typos, mis-assignment, and non-assignment. In GenBank, this problem concerns roughly 18% of records, an alarming number that should worry the biocuration community.

To address this problem of automatic sequence type classification, we propose the use of literature associated to sequence records as an external source of knowledge that can be leveraged for the classification task. We define a set of literature-based features and train a machine learning algorithm to classify a record into one of six primary sequence types. The main intuition behind using the literature for this task is that sequences appear to be discussed differently in scientific articles, depending on their type. The experiments we have conducted on the PubMed Central collection show that the literature is indeed an effective way to address this problem of sequence type classification. Our classification method reached an accuracy of 92.7%, and substantially outperformed two baseline approaches used for comparison.

KEYWORDS

Data Analysis; Data Quality; Biological Databases; Data Cleansing.

1 INTRODUCTION

Bioinformatics sequence databases such as GenBank or UniProt contain large numbers of nucleic acid sequences and protein sequences. In the context of uncurated databases, records are primarily defined, uploaded, and annotated exclusively by the submitters themselves without any particular quality control mechanism. Given this large amount of data and the nature of the submission process, the records suffer from a large range of data quality issues [8] including errors, discrepancies, redundancies, ambiguities, and incompleteness. These quality issues can seriously hamper the efficacy of data mining, and machine learning algorithms.

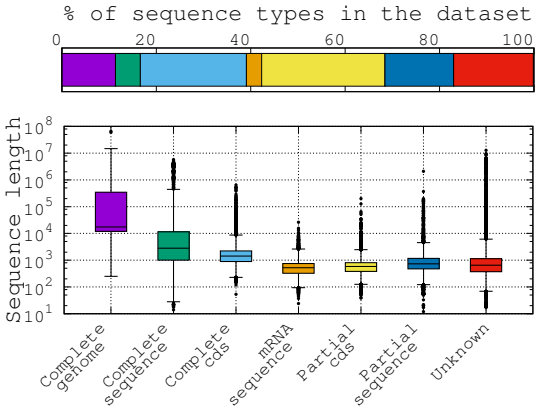

Figure 1: Sequence type statistics.

The published scientific literature has been recently considered as a source of evidence to help biocurators to detect faulty and suspicious records that are inconsistent with the literature [1, 2]. In this work, we expand on this approach to explore the use of the literature as the basis of a classifier to help biocurators in identifying and assigning correct sequence types. Indeed, a sequence type may be assigned to each record in a biological sequence database to provide information about the nature of the sequence represented in that record, and can be used in search to filter out irrelevant sequences (e.g., search for genome of …, search for coding sequences related to gene …, etc.). However, in GenBank the sequence type attribute is a non-mandatory free-text field, and thus it is subject to many errors including typos, mis-assignment, and non-assignment. Moreover, as there is no clear nomenclature for the sequence type attribute, it is hard to extract the sequence type from the free text field. Focusing on GenBank, we consider in this paper six sequence types for classification: (i) complete genome, (ii) complete sequence, (iii) complete coding sequence (cds), (iv) mRNA sequence, (v) partial cds, and (vi) partial sequence.

To get an overview of the current quality of the sequence type attribute in GenBank, we refer to Figure 1. The figure shows both the overall proportion of sequence per type (top figure) and the distribution of sequence length per type using box-plots (bottom figure) for the GenBank dataset discussed in Section 4. There are four notable trends here: (i) First, based on the top figure, roughly 18% of records have unknown type, which from a data quality point of view can be considered as a serious problem as many records will

CIKM'17, November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
DOI: https://doi.org/10.1145/3132847.3133051

not be able to be retrieved or filtered by type. Here, the sequence type may have been written with spelling errors or other such small mistakes (see record with accession number AP013068[1]), or may have been omitted entirely (see record with accession number AP011615[2]). (ii) Second, given the sequence length distribution, it is clear that complete genomes/sequences tend to have longer sequence length than partial sequences. (iii) Third, looking at the sequence length distribution for sequences with unknown type, we notice that this quality issue appears to affect all sequences regardless of length. (iv) Fourth, given a sequence length, it is not absolutely possible to guess the sequence type. For example, viroids, which are considered as the smallest infectious pathogens have a complete genome of 249 bp.[3]

Given these observations, we propose in this paper an automatic method for sequence type classification. Our approach is primarily grounded in analysis of the literature, based on a previous study that shows the role of literature consistency in flagging possibly erroneous records in GenBank [2]. The intuition is that different sequence types are discussed differently in articles. Using the list of published research articles associated with each biological sequence record with known sequence type, and a set of features that capture different aspects of those articles, we train a machine learning algorithm that will classify a GenBank record into one of the six sequence types defined above, based on its associated literature. To the best of our knowledge, this work is the first attempt that proposes to recognise biological sequence types using the published literature.

2 BACKGROUND AND PRELIMINARIES

Here, we first describe the structure of a sequence record in GenBank and then specifically define the problem we study.

2.1 GenBank sequence record structure

The format of a sequence record can be regarded as having three parts: the header, which contains the information that applies to the whole record; the features, which are the annotations on the sequence; and the sequence itself. The header section is composed of several fields: (i) *LOCUS* field: contains a number of different data elements, including locus name, sequence length, molecule type, and modification date; (ii) *DEFINITION* field: a brief description of sequence or sequence's function, where usually the sequence type is given in a free text entry; (iii) *ACCESSION* field: a unique identifier for the record; (iv) *SOURCE* field: gives information about the sequence's organism; (v) *REFERENCE* field: lists a set of publications by the authors of the sequence that discuss the data reported in the record.

It is clear that the header part represents a rich source of information. Hence, based on the fact that articles discuss the data reported in the records, and that the record definition provides a summary of the major information reported in that record, we will primarily focus on the match between record definitions and the associated literature, in order to predict the record sequence.

[1]https://www.ncbi.nlm.nih.gov/nuccore/AP013068
[2]https://www.ncbi.nlm.nih.gov/nuccore/AP011615
[3]https://www.ncbi.nlm.nih.gov/nuccore/KC581915

2.2 Research problem statement

We define the problem we study in this paper as follows:

Given (i) a collection of documents which represents the domain literature knowledge $D =< d_1, d_2, ..., d_n >$; (ii) a set of annotated records $R =< (r_1, y_1), (r_2, y_2), ..., (r_m, y_m) >$, where $y_m \in \{$*complete genome*, *complete sequence*, *complete cds*, *mRNA sequence*, *partial cds*, *partial sequence*$\}$; and (iii) the set of documents associated to each record $D_R =< D_{R_1}, D_{R_2}, ..., D_{R_m} >$, the problem we study is: for a new record r and its set of associated documents D_R, we aim to predict the sequence type y of r.

3 LEARNING SEQUENCE TYPE

Our objective is to classify a record according to one of the six sequence types defined in the previous section. For that purpose, we define and use a set of literature-based features that estimate how a sequence record is discussed in the literature. We consider three literature-based feature types for each sequence record r.

First, for each sequence record r, we consider a set of features based on the similarity of its definition and its associated set of documents D_R. These similarity measures are [10]: matching, overlap, Jaccard, Dice, cosine, mutual information (MI), and Okapi BM25 [11]. We also used various IR similarity ranking functions including the sum of TF-IDF scores (SumTFIDF), the Lucene vector-space model score (LuceneVSM),[4] the BM25 score [11], language model scores based on (i) the Jelinek-Mercer smoothing (LMJelinekMercer) [13] and on (ii) a Bayesian smoothing using Dirichlet priors (LMDirichlet) [13], and an information-based score (IBSimilarity) [4]. These similarities are computed separately for each of four different document fields {title, abstract, body, all document}.

Second, for each r, we consider a set of frequency based features that estimate how well r is discussed into its associated set of documents D_R. These frequency-based features include *term frequency* (TF), *inverse document frequency* (IDF), and the TF-IDF score. Note that these scores are derived from term level statistics. Hence, for a sequence record r, we calculate aggregated values using the sum, standard deviation, minimum, maximum, arithmetic mean, geometric mean, harmonic mean, and coefficient of variation of the IDFs of constituent terms, again considering separately the four different fields of a document.

Third, we also consider a set of information retrieval-based features, specifically those related to query quality prediction. In this case, we consider a record definition as a query, and its associated documents are considered as the set of relevant documents. Here, we consider: query clarity [6], simplified clarity score [7], similarity of collection–query score [14], inverse collection term frequency features [9], and query scope [7]. These information retrieval-based are, as for the other features, computed while considering separately the four document fields.

Therefore, in total, we consider 171 literature-based features that characterize how a record is discussed in the literature, and in particular in its associated documents. The features that we have described above are explained in more detail elsewhere [1].

Given as input a set of literature-based features for each record r, our goal is to combine these inputs to produce a value y indicating

[4]https://lucene.apache.org/core/6_1_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html

the sequence type of that record. To accomplish this, we used the Support Vector Machines (SVM) classification algorithm [5], one of the most widely-used and effective classification algorithms.

Each record r is represented by its vector of k quality indicators $x_r = [x_{r1}, x_{r2}, ..., x_{rk}]$ and its associated label $y_m \in \{$*complete genome*, *complete sequence*, *complete cds*, *mRNA sequence*, *partial cds*, *partial sequence*$\}$. We used the SVM implementation available in the LibSVM package [3]. Both Linear and Radial Basis Function kernels were considered in our experiments. The regularization parameter C (the trade-off between training error and margin) and the gamma parameter of the radial basis function kernel were selected from a search within the discrete sets $\{10^{-5}, 10^{-3}, ..., 10^{13}, 10^{15}\}$ and $\{10^{-15}, 10^{-13}, ..., 10^{1}, 10^{3}\}$ respectively, using 10-fold cross validation. Although the differences were not substantial, experiments with the best radial basis function kernel parameters performed slightly better than the best linear kernel parameters for the majority of the validation experiments. Thus, all presented results were obtained using an radial basis function kernel, with C set to 10^{15} and gamma set to 10^{-11}.

4 DATA DESCRIPTION

Articles: We used the PubMed Central Open Access collection[5] (OA), which is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine. The release of PMC OA we used contains roughly 1.13 million articles, which are provided in an XML format with specific fields corresponding to each section or subsection in the article. We used the Lucene IR System[6] to index the collection, with the default settings for stemming and English stopword removal. We defined a list of biomedical keywords, which should not be stemmed or considered as stop-words, such as the protein names *"THE"* and *"Is"*. Each section of an article (title, abstract, body) is indexed separately, so that different sections can be used and queried separately to compute the quality features.

Sequences: We work with the GenBank nucleotide database, but limit the sequence database records we work with to those that are cited by the PMC OA article collection. Specifically, we used a regular expression to extract GenBank accession numbers mentioned in the PMC OA articles, thereby identifying literature that refers to at least one GenBank identifier. This resulted in a list of 733,779 putative accession numbers. Of these, 494,142 were valid GenBank nucleotide records that we were able to download using the e-utilities API ([12]).[7] Among the valid records, only 162,913 records also cite the corresponding articles (as determined by matching their titles). This process gave us a list of 162,913 pairs of record accession numbers and PMC article identifiers, which cite each other. Note that for the 331,229 records that we have put aside, each record cites an article; however, we do not have access to all articles through PMC OA. In order to avoid overfitting for the model where records that belong to the same articles learn from each other, we have decided to randomly choose only one record per article to train the model. Hence, we reduced the set of record we use in our experiments to 15,133 records.

[5] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ The version used was downloaded on October 2015.
[6] http://lucene.apache.org/
[7] The sequences were downloaded on October 2015.

Table 1: Detailed accuracy by class.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
(i)	0.967	0.005	0.970	0.967	0.969
(ii)	0.800	0.007	0.881	0.800	0.839
(iii)	0.919	0.032	0.913	0.919	0.916
(iv)	0.896	0.003	0.923	0.896	0.910
(v)	0.936	0.035	0.925	0.936	0.930
(vi)	0.945	0.014	0.934	0.945	0.939
Avg.	**0.927**	**0.023**	**0.927**	**0.927**	**0.927**

Labels: As stated previously, the sequence type of a record is often given in the *definition* field as a free text entry. For example, the definition of the record with accession number KC581915 is: "*Coleus blumei viroid 1 isolate BJ-1-1, complete genome*". Hence, we used a simple string matching mechanism to split a record definition into two parts: (1) the sequence type that is used to label the dataset, and (2) the remainder of the record definition. The sequence type information is removed from the definition text that is used to compute the features in order to simulate the scenario of missing or erroneous sequence type in the record.

After processing, among the 15,133 records, we have (i) 2,069 complete genome, (ii) 961 complete sequence, (iii) 4,097 complete cds, (iv) 589 mRNA sequence, (v) 4,765 partial cds, and (vi) 2,652 partial sequence. The dataset we built can therefore be considered to be fairly balanced.

5 EMPIRICAL EVALUATION

We now report and discuss the main results of the experimental evaluation, considering both the effectiveness of the method and our interpretation of which features are valuable in classification.

5.1 Performance Analysis

Table 1 shows the accuracy of our classifier broken down by the six classes in the data. The last row of the table shows the overall accuracy. The overall accuracy of 92.7% shows the effectiveness of the literature-based features we have described to discriminate between sequence types. We note that the best accuracy (96.9%) is obtained when classifying *complete genomes*, suggesting that these sequences are discussed in a particular way in the literature. Indeed, a specific research article is usually devoted to the description of each *complete genome*. The results obtained at this stage confirm our initial assumption that sequences are discussed differently in the literature based on their types.

In order to show the effectiveness of the method we described, we provide a comparison with two baseline methods:

- SVM Record-based Features (RBF): SVM classifier trained with only record-based features, derived from the records themselves, including the record popularity, organism popularity, number of coding regions, definition length, etc.
- RandomTree: Random Tree classifier trained using sequence length only.

Table 2: Performance comparison.

Algorithm	Features	Precision	Recall	F-Measure
SVM LBF	Literature	0.927	0.927	**0.927**
SVM RBF	Record	0.502	0.561	0.506
RandomTree	Seq. length	0.469	0.496	0.473

The comparison results are presented in Table 2, where we refer to our approach as SVM Literature-Based Features (LBF). The results show that our approach outperforms the two baselines, with an improvement of roughly 83% over the best baseline. Moreover, the results show that the literature-based features are highly effective for detecting sequence types, compared to other features like sequence length and record internal features. We can conclude that the literature is an effective source of evidence to help biocurators to automatically determine the appropriate record sequence type.

5.2 Feature Analysis

We now analyze the informativeness of our defined features in Section 3 and consider their effect on learning targeted classifiers. We use Mutual Information (MI) as our primary metric for feature evaluation, where higher values for MI indicate more informative features for the given topic.

We provide the mean Mutual Information values for each feature across different topics in Figure 2. The last column in Figure 2 shows the Mutual Information over all classes. We observe that across all classes, query quality features are the most informative features. Looking at the overall MI values, the order of feature types from most to least informative is the following: query quality features, frequency features, and similarity features.

The informativeness of such query quality features indicates that they are more sophisticated and more elaborated than the other feature types. The results indicate that an information retrieval approach that considers a record definition as a query and the set of documents related to that record as the set of relevant documents for the query is a meaningful approach for modelling the sequence type classification task.

Finally, we have identified the top 5 most informative features as: inverse document frequency, inverse collection term frequency score, similarity of collection–query score, query scope score, and clarity score. Also, comparing art sections, features were more informative while computed over titles, followed by abstracts and then article bodies. This suggests that short sections are more informative than long sections; they are probably less noisy.

6 CONCLUSION

We have proposed in this paper a new method to automatically classify sequence types for biological sequence records, by using features derived from the published literature associated to those sequence records. The evaluation we have carried out shows that the literature is a better source of information to address the problem of sequence classification, than information derived from the records themselves, including the sequence length. Hence, the main outcome of this paper is that we have shown that the literature is a meaningful source of evidence for this task. The classification of

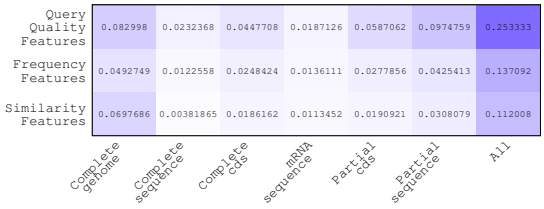

Figure 2: Mutual Information values for each class per feature type, and overall.

sequence records in this way can help biocurators to detect and correct errors, and will specifically result in higher reliability of sequence type information in sequence databases.

Future work includes exploring the use of literature to address other biocuration issues such as duplicate detection. We also plan to consider the validity of the approach for new records that may have less associated literature, as compared with records with abundant referring literature.

Acknowledgements: This work is supported by the Australian Research Council through a Discovery Project grant, DP150101550.

REFERENCES

[1] M. R. Bouadjenek, K. Verspoor, and J. Zobel. Automated detection of records in biological sequence databases that are inconsistent with the literature. *Journal of Biomedical Informatics*, 71:229–240, 2017.
[2] M. R. Bouadjenek, K. Verspoor, and J. Zobel. Literature consistency of bioinformatics sequence databases is effective for assessing record quality. *Database*, 2017(1):bax021, 2017.
[3] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.
[4] S. Clinchant and E. Gaussier. Information-based models for ad hoc ir. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 234–241, New York, NY, USA, 2010. ACM.
[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
[6] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 299–306, New York, NY, USA, 2002. ACM.
[7] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *String Processing and Information Retrieval: 11th International Conference, SPIRE 2004, Padova, Italy, October 5-8, 2004. Proceedings*, pages 43–54, Berlin, Heidelberg, 2004. Springer.
[8] J. L. Y. Koh, M. L. Lee, and V. Brusic. A classification of biological data artifacts. In *Workshop on Database Issues in Biological Databases*, pages 53–57, 2005.
[9] G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 564–571, New York, NY, USA, 2009. ACM.
[10] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 641–650, New York, NY, USA, 2009. ACM.
[11] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-2. In *TREC*, pages 21–34, 1993.
[12] E. Sayers. E-utilities quick start. entrez programming utilities help. Technical report, 2010.
[13] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 334–342, New York, NY, USA, 2001. ACM.
[14] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, pages 52–64, Berlin, Heidelberg, 2008. Springer.