Identification of nucleobase sequences with text processing methods

James Martin
Computer Science Student
University of Southampton
Southampton, England

Abstract—**Many similarities exist between nucleobase/amino acid sequences that form genes/proteins and word sequences that form unstructured text. This paper explores the relationship between the two and whether natural language processing methods can be used for sequence identification. To achieve this, three programs were developed that use the bag of words, frequency analysis, and brute force approaches, and they were tested to evaluate their efficacy. This paper found bag of words to be very effective at predicting the species of unseen gene sequences and concluded that the metaphor of genes as unstructured text is useful.**

Keywords—Computational Biology; Natural Language Processing; Sequence Matching; Bag of Words

I. Introduction

In computational biology working with genes requires the analysis and computation of massive strings of characters, each representing a nucleobase base – a process with many parallels to working with natural language. Due to computational biology as a field developing in the 1970s [1], later than natural language processing in the 1950s [2], many methods used in the processing of genomic data originated from the processing of text. This paper explores the relationship between sequence processing and natural language processing, and the techniques shared between them.

A. *Research Objectives*

To this end, three research objectives can be defined:

1) To find similarities between text processing and computing biological data.

2) To apply different natural language processing methods to perform nucleobase sequence identification.

3) To test the effectiveness of each method.

II. Background

Natural language processing methods are used frequently in computational biology, with applications such as biomedical knowledge graphs [3] and sequence and structure alignment [4]. This is due to the similarities between gene/protein sequences and natural language; the main ones being that both have alphabets (ATCG for genes, A-Z for text), the order of the letters dictates the meaning and form words (codons for genes) [5], and both are read linearly – this points to the high applicability of natural language processing algorithms to biological sequence identification.

A. *Natural Language Processing (NLP)*

NLP is the processing of unstructured text and is said to be comprised of three separate ideas: the retrieval of information, semantic analysis, and the extraction of information [6], though this paper will focus on the semantic analysis aspect, which involves parsing and classifying text to identify it or determine meaning. Many techniques exist to classify unstructured text, such as Bag of Words, Naïve Bayes, and Support Vector Machines [7] - this paper focuses on the bag of words technique.

B. *Bag of Words (BOW)*

A commonly used model for related documents and their contents is bag of words as it can represent information efficiently and determine similarity between documents without needing to understand them. This involves, for each document, counting each normalised word or phrase and creating a sparse vector to store the counts, then taking the cosine similarity between two document's vectors to determine the similarity between their subject matters [8]. The model can be used for classification in areas other than natural language, for example the bag of visual words technique involves extracting important features from images and calculating the similarity between the resultant feature vectors to determine the closest classification. If genes are able to be considered as documents and sections of nucleobase sequences can be words or phrases, then the bag of words model could be used to classify genes.

C. *Sequence Identification and Alignment*

This is the identifying of similar regions of nucleobase or amino acid sequences to match genes/proteins or determine evolutionary similarity. Many algorithms exist to perform this task, including dynamic programming methods such as Needleman-Wunsch or Smith-Waterman, and heuristic methods such as BLAST [9]. The issue with a lot of these methods is that they take an excessive amount of time to run, especially with large databases to search.

D. *Summary*

There are many similarities between nucleobase sequences and unstructured text, which has led to the use of many NLP methods in bioinformatics applications. The literature reviewed points to NLP methods being already well developed and suitable for processes such as nucleobase sequence matching.

III. Methodology

To apply NLP methods to nucleobase sequence identification, three programs were created; two using NLP methods and one using brute force searching to act as a baseline for the others. The two NLP methods selected were bag of words and word frequency analysis, and all programs were developed using Python. The programs take a base sequence file as input and output the closest matching gene from the training data, and whether it is from a human or rat chromosome.

A. *Implementation*

- The implementation of the bag of words program is as follows: first the training file data is read and converted to vectors such that only 10-long base sequences are taken and converted after the appearance of double-repeated bases (such as ‘AA’) – this is preferred over taking random selections to achieve shift-invariance. These vectors are then clustered using k-means to quantise a set of vectors, forming a vocabulary, and histograms are created for each training file recording a count of vocabulary words present. To identify new test data, it is first read, vectors are extracted, quantised, and turned into a histogram, then the cosine similarity between the test file histogram and each train file histogram is calculated – the best similarity giving the closest gene match and species identification.

- The implementation of the word frequency analysis differs in that the vocabulary is predefined to be permutations of 4-base long sequences, and histograms record the presence of each in every training and test file. The histograms are again compared using cosine similarity.

- The third program stores all training data in arrays and brute-force searches for the best match at all possible sequence start points, as long as the start of the sub-sequence appears promising.

All three programs underwent parameter optimisation, and the two NLP method programs went through many iterations of feature extraction methods/vocabularies to maximise efficacy.

B. *Train and Test Data*

Training data consisted of FASTA-format base sequence files representing genes involved in cell mitosis, including: SGO1, NUMA1, NEK2, and seven others. Two copies of each gene were included in the training set, from both the human and brown rat chromosomes. This data was sourced from the NIH GenBank database. Test data included both small snippets of genes from the training set (seen), and complete additional genes not included in the training set (unseen, including: CDK1, FST, NEK4, and 5 others) – this is to test both gene identification and species identification. All testing data was also from genes involved with mitosis. In total 20 nucleobase sequence files were used for training, and 36 were used for testing.

To test each program all test files were used, and the programs were run 20 times to get an average accuracy – measured by the percent of test data correctly categorised.

IV. Results

TABLE I. AVERAGE ACCURACY AND PERFORMANCE OF METHODS TESTED

Methods	**Average Accuracy (%)**			**Time To Run (S)**	**Trained Size (KB)**
	Seen (Species)	***Seen (Gene)***	***Unseen (Species)***		
Bag of Words	96.75	85.50	85.63	1.05[a] 2.74[b]	65
Frequency Analysis	70.00	35.00	75.00	0.30[a] 0.76[b]	3
Brute Force	100.00	100.00	62.50	136.07	956

[a] Excluding training.

[b] Including training.

The trained size heading refers to the total size in kilobytes of all histograms and arrays produced after training, that are used for classifying new data.

The most accurate program at identifying the species and gene of the seen test data was the brute force method as expected, as it checks exactly every possible sequence match. It did however have by far the worst time and size performance - 50x and 15x the second worse scores respectively. In use cases where extreme accuracy is required this is still a valid method, but it would likely be used in combination with other heuristic methods to increase performance while remaining accurate. The one area this method fails in is dealing with unseen data, as 62% accuracy is too low to be called useful with such a small train and test sample size.

The bag of words method far outperformed the others in the unseen data tests, whilst also having a 96% accuracy on the seen data tests. This make the BOW method seem very promising for fast and fairly accurate sequence identification, however more testing would be needed with larger datasets to fully gauge its usefulness. It should be noted that training the BOW model took up most of the processing time, meaning the short amount of time it takes to classify new data with precomputed training histograms present a great benefit over the brute force solution with little loss of accuracy when predicting species.

The frequency analysis method achieved better than the brute force method with the unseen data tests, but overall had low average accuracies. It was the method with the lowest time and size costs however, meaning if the choice of vocabulary can be improved enough, it could be beneficial in resource-limited situations.

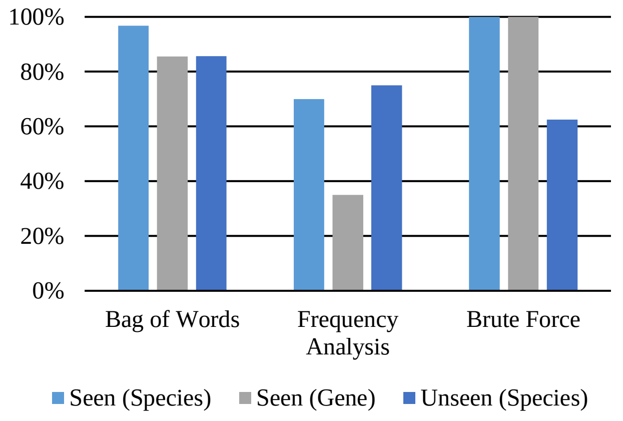

Fig. 1. Average accuracy of methods tested as a multi-bar graph.

V. Evaluation

Not every NLP method translates perfectly to computational biology, but with the proper modification they can be invaluable when working with sequences. The bag of words method shows great promise in fast sequence prediction, but it is unknown how well this scales with larger training data sets. The frequency analysis appears far less useful but would need further experimentation with vocabularies to know with certainty.

VI. Conclusion

A. Summary

The first research objective was completed in the background chapter where similarities between text processing and biological computing were identified, the second objective was completed by developing the three programs using two NLP methods, and the third objective was completed by testing the three programs using gene data from the GenBank database. The tests performed with the programs developed shows that NLP methods can be used to perform bioinformatics tasks, which support the comparison between text/words and sequence/codons.

B. Future Work

Due to time constraints this project had to be limited in its scope, and there are multiple ways in which it could be expanded to determine more about the NLP methods tested. Firstly, the choice of vocabulary for the frequency analysis method can be explored further; due to the infinite number of possible choices the final program run for the tests could have a great potential for improvement. Secondly, tests could be performed with larger training and testing data sets including more genes to get more reliable results and further optimise the programs – it is possible that the great performance of the bag of words method was only due to the small sample size used in testing, so more data is needed to evaluate this. Finally, testing and training data only included genes that are involved with cell mitosis, so further work could be done to test the methods with data sets that include a greater variety of gene types.

References

[1] P. Hogeweg, “The roots of bioinformatics in theoretical biology,” *PLoS computational biology*, vol. 7, issue. 3, March 2011.

[2] E. D. Liddy, “Natural Language Processing,” *Encyclopedia of Library and Information Science, 2nd Ed*, NY, Marcel Decker Inc, 2001.

[3] K. Hänsel, S. N. Dudgeon, K. H. Cheung, T. J. S. Durant, and W. L. Schulz, “From Data to Wisdom: Biomedical Knowledge Graphs for Real-World Data Insights,” *Journal of medical systems*, vol. 47, issue. 1, p. 65, May 2023.

[4] Z. Zeng, H. Shi, Y. Wu, and Z. Hong, "Survey of Natural Language Processing Techniques in Bioinformatics," *Computational and Mathematical Methods in Medicine*, vol. 2015, October 2015.

[5] D. Ofer, N. Brandes, M. Linial, “The language of proteins: NLP, machine learning & protein sequences,” *Computational and Structural Biotechnical Journal*, vol. 19, pp. 1750–1758, March 2021.

[6] M. Yandell, and W. H. Majoros, “Genomics and natural language processing,” *Nature Reviews Genetics*, vol. 3, pp. 601–610, August 2002.

[7] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text Classification Algorithms: A Survey,” *Information*, vol. 3, issue. 4, p. 150, April 2019.

[8] D. Gunawan, C. A. Sembiring, and M. A. Budiman, “The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents,” *Journal of Physics: Conference Series*, vol. 978, March 2018.

[9] E. S. Donkor, N. T. K. D. Dayie, and T. K. Adiku, “Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA),” *Journal of Bioinformatics and Sequence Analysis*, vol. 6, issue 1, pp. 1–6, April 2014.