Identification of Ancient Greek Papyrus Fragments Using Genetic Sequence Alignment Algorithms

Alex C. Williams, John F. Wallin, Hyrum D. Carroll
Middle Tennessee State University

Introduction

Papyrologists analyze, transcribe, and edit papyrus fragments in order to enrich modern lives by better understanding the linguistics, culture, and literature of the ancient world. One of their common tasks is to take an anonymous fragment and to identify a matching known manuscript. This is especially challenging when the fragments are damaged and contain only limited information (*e.g.*, due to deterioration). In the last 100 years, only about 10% of the well-over 500,000 fragments recovered from the Egyptian village of Oxyrhynchus have been edited [4]. We do not know what new ancient texts might be found and what can be learned from them, but using current methods of identification this process will take in excess of 1000 years.

A similar problem exists in computational biology. When a new, unknown gene is found, biologists compare the genetic sequence of the new gene to sequences of existing, known genes in order to identify and classify the new gene. This process is called genetic sequence alignment. Previous studies have shown that computational biology algorithms and solutions can be usefully applied to problems in other disciplines [2]. In order to accelerate the tedious process of manual papyri identification, we introduce Greek-BLAST, which utilizes genetic sequence alignment algorithms as a method for identification of Ancient Greek text fragments. Our preliminary results using naive methods have already demonstrated the applicability of this approach.

Objectives

The purpose of this study is to produce a computational tool capable of automatically aligning unidentified Greek papyrus fragments to known manuscript counterparts in order to accelerate the manual process of papyrus identification. A secondary purpose of this study is to understand the necessary steps in re-tailoring domain-specific sequence alignment algorithms (*i.e.*, BLAST) to alternative domains.

Methods

Development of Greek-BLAST

The Basic Local Alignment Search Tool (BLAST) is a computational tool for aligning genetic sequences [3]. Greek-BLAST deviates from the traditional BLAST algorithm by handling fragments of Greek text instead of genetic sequences. It is based on version 2.2.27 of NCBI's BLAST. Figure 1 displays the BLAST algorithm.

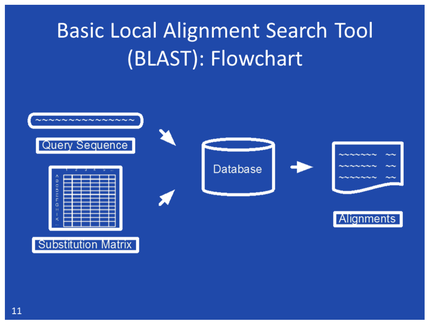

Fig. 1 The BLAST algorithm.

Fig. 2 A papyrus fragment from the Oxyrhynchus collection..

Calculating a New Substitution Matrix

In BLAST, the substitution matrix stores statistical values for representing the likelihood of one character in an set of letters being replaced with another character in the same set. As the identification performance of BLAST is directly influenced by the substitution matrix being used for detecting alignments, substitution matrices have been studied rigorously to improve alignment accuracy between genetic sequences [5]. Presently, Greek-BLAST naively operates using an identity substitution matrix.

A new substitution matrix that represents the true likelihood of character replacements or misidentifications in the transcription process will dramatically increase identification performance. A new substitution matrix will be calculated using a log-odds approach that empirically represents the letter statistics of the Greek language. By creating this new matrix, the likelihood of being able to correct match and identify an unknown fragment will greatly increase.

Results

Preliminary Results and Future Datasets

Using a database of 6,600 known manuscripts, Greek-BLAST was able to identify and match 18 fragments to a known manuscript. To strengthen the identification performance of Greek-BLAST, key variables (i.e., insertion rate, deletion rate, fragment length, error rate) will be studied and identified using simulated Greek fragments.

In addition to simulated fragments, we will also use the crowdsourced transcriptions of real, unidentified fragments from the University of Oxford's Ancient Lives project [1]. (See Fig. 3) These transcriptions will be used as input to our final implementation of Greek-BLAST to evaluate the identification performance of the algorithm.

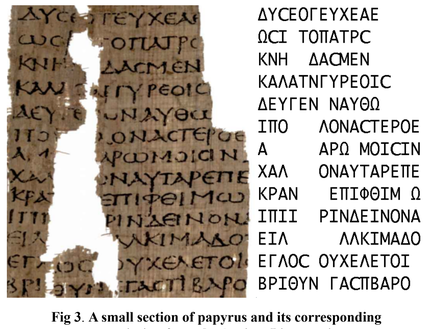

Fig 3. A small section of papyrus and its corresponding transcription from the Ancient Lives project.

Conclusions

In this study, we observed a deficiency in the current rate of Ancient Greek papyri identification. Papyrologists try to manually match an anonymous, unknown Ancient Greek papyrus fragment to a known Ancient Greek full-text manuscript. This process can take days, weeks, or even years to match a single papyrus fragment. In order to hasten the tedious process of manual identification, we introduced a new methodology that aims to leverage genetic sequence alignment algorithms for Ancient Greek papyrus identification.

Future work includes applying this methodology to complete the development of Greek-BLAST, a BLAST variant adapted to the Greek text domain. Such a tool would allow papyrologists to leverage modern computational methods in order to dramatically increase the rate of identification for Ancient Greek papyrus fragments. Greek-BLAST will offer a new form of non-contextual text identification that is not currently available for the Greek language.

Acknowledgments

We would like to acknowledge Nita Krevans, Dirk Obbink, Marco Perale, Phillip Sellew, and Trevor Wennblom for their contributions to this project. Additionally, we would also like to acknowledge Egyptian Exploration Society for access to the Oxyrhynchus papyri collection.

Alex Williams and Hyrum Carroll were supported in part by a Faculty Research and Creative Activity Award grant (2-21659) from Middle Tennessee State University.

References

[1] Ancient Lives. https://ancientlives.org.
[2] Abouelhoda, M. and Ghanem, M. String Mining in Bioinformatics. In Scientic Data Mining and Knowledge Discovery, pages 207{247. Springer, 2010.
[3] Altschul, Stephen F., et al. "Basic local alignment search tool." *Journal of molecularbiology* 215.3 (1990): 403-410.
[4] Bowman, A.K., Coles, R.A., Gonis, N., Obbink, D., and Parsons, P.J.. Oxyrhynchus: a City and its Texts, volume 93. Egypt Exploration Society, 2007.
[5] **Chiaromonte, F., V. B. Yap, and W. Miller.** "Scoring pairwise genomic sequence alignments." *Pacific Symposium on Biocomputing*. Vol. 7. 2001.