COMPUTATIONALLY ACCELERATED PAPYROLOGY

By

Alex C. Williams

A thesis submitted in partial fulfillment

of the requirements for the degree of

MASTER OF SCIENCE

in

Computer Science

Middle Tennessee State University

May 2015

Thesis Committee:

Dr. Hyrum D. Carroll

Dr. John F. Wallin

Dr. Cen Li

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Hyrum Carroll, for his wonderful mentorship and his contributions to my growth as both a professional researcher and computer scientist. I would like to also thank Dr. John Wallin for encouraging me to be think creatively and introducing me to the exciting area of citizen science. I would also like to acknowledge the faculty within the Department of Computer Science, particularly Dr. Chrisila Pettey, who have graciously offered their support and encouragement in completing this thesis.

I would like to thank my fellow graduate students, Anthony Mills, Anthony Davis, Matt Houglum, Robert Reaves, Matt Wang, and Nathan Reale, for their advice and patience in listening to me ramble about my thesis.

I would like to extend the largest of thanks to the University of Oxford's papyrology staff, which includes James Brusuelas, Chiara Meccariello, Gertjan Verhasselt, Maroula Salemenou, and Annapaola Santarsiero, for their assistance in developing better software for papyrology and contributing to the strength of this thesis.

I would like to acknowledge and thank the Egyptian Exploration Society for access to the dataset used in this thesis.

Lastly, I would like to thank Roya for her encouragement, love, and support throughout the completion of this thesis. Most importantly, I would like to thank both my parents, Mike and Tammy, to whom this thesis is dedicated, for inspiring me to pursue my dreams and bring positive change to the world through my work.

ABSTRACT

Papyrologists transcribe and identify papyrus fragments in order to enrich modern lives by better understanding the linguistics, culture, and literature of the ancient world. In practice, these tasks are extremely challenging and slow due the limited amount of information preserved in each papyrus fragment (*i.e.*, due to deterioration). For example, since their discovery in the late 19th century, only 10% of the more than 500,000 fragments in the Oxyrhynchus papyri collection has been given preliminary identifications.

This thesis presents two computational approaches for accelerating papyrus transcription and identification. The first approach is a computational pipeline that aggregates millions of crowdsourced letter classifications into transcriptions of papyrus fragments. The second approach leverages genetic sequence alignment algorithms to rapidly identify damaged papyrus fragments to known papyrus manuscripts. These approaches greatly improve upon the current state-of-the-art techniques and set a new standard for leveraging computation to the transcription and identification of ancient texts.

PREFACE

Chapter 3 is a version of the conference paper, *A Computational Pipeline for Crowdsourced Transcriptions of Ancient Greek Papyrus Fragments*[1]. The University of Oxford's Zooniverse team and Ancient Lives project, which includes Chris Lintott, John Wallin, and James Brusuelas, created the Ancient Lives interface and provided the crowdsourced data used in this study. Collaborators from the University of Minnesota, including Lucy Fortson, Anne-Francoise Lamblin, and Haoyu Yu, contributed greatly to the initial stages of project development. I designed and developed the pre-processing stage for separating individual click-data for each fragment into distinct files and determining an image's most probable orientation. John Wallin and Haoyu Yu designed and developed the kernel-based approach for aggregating individual user letter identifications. John Wallin also designed and developed the stepwise approach, which I have made very few modifications. I developed the approach used in Stage 2 of the pipeline which calculates line sequences. John Wallin designed and developed an additional version of Stage 2 that was referential in the design of my approach. I developed a high-level wrapper for both executing and evaluating the pipeline. I also developed visualization software used to overlay consensus letters and letter-groups on an image (see Figure 8). Marco Perale manually performed a preliminary evaluation of the kernel-based approach. Marco Perale, James Brusuelas, and Dirk Obbink provided letter identifications from the published Oxyrhynchus volumes that were used in evaluating both stages of the pipeline. I conceived the study's two-part evaluation. Additionally, I performed the analysis of the results for each stage and wrote the software for creating Figure 9. Additionally, I wrote the manuscript for the published paper. John Wallin was the supervisory author of this project and has been involved heavily throughout the project in concept formation and continuous development.

[1] Williams, A. C., Wallin, J. F., Yu, H., Perale, M., Carroll, H. D., Lamblin, A. F., Fortson, L., Obbink. D, Lintott, C.J., & Brusuelas, J. H. (2014, October). A Computational Pipeline for Crowdsourced Transcriptions of Ancient Greek Papyrus Fragments. In Big Data (Big Data), 2014 IEEE International Conference on (pp. 100-105). IEEE.

Chapter 4 is a version of the conference paper, *Identification of Ancient Greek Papyrus Fragments Using Genetic Sequence Alignment Algorithms*[2]. Hyrum Carroll and John Wallin initiated the study by implementing an identity substitution matrix in BLAST, converting a proprietary database of Greek manuscripts into a FASTA format, and creating the encoding scheme between amino acids and Greek letters. Collaborators from the University of Minnesota, including Anne-Francoise Lamblin and Haoyu Yu, contributed additional preliminary work with investigating the applicability of genetic sequence alignment. I investigated applicable methodologies for calculating a new substitution and selected the BLOSUM methodology. Both Hyrum Carroll and I conceived how to leverage the BLOSUM methodology. I developed software for (a) calculating background frequencies of each letter in the Greek alphabet, (b) calculating misidentification ratios for each letter pair in the Ancient Lives crowdsourced data, and (c) calculating a new substitution matrix with the calculated letter background frequencies and misidentification ratios. I implemented the calculated matrix into BLAST. Hyrum Carroll provided software that I heavily modified to create the 14,100 simulated fragments used to evaluate the new matrix. I conceived and performed the study's evaluation. I independently analyzed the results of the evaluation and both authored and presented the published paper. Hyrum Carroll and John Wallin are the supervisory authors for this study and were involved to a great degree during project conception.

The work in both Chapters 3 and 4 would not have been possible without the millions of contributions made by the thousands of generous volunteers from the Zooniverse project. To each volunteer, I would like to extend a special thanks for allotting their valuable time to making each of these projects possible.

[2]Williams, A. C., Carroll, H. D., Wallin, J. F., Brusuelas, J., Fortson, L., Lamblin, A. F., & Yu, H. (2014, October). Identification of Ancient Greek Papyrus Fragments Using Genetic Sequence Alignment Algorithms. In e-Science (e-Science), 2014 IEEE 10th International Conference on (Vol. 2, pp. 5-10). IEEE.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER I. INTRODUCTION 1

Papyrology 2

Thesis Statement 4

Thesis Outline 6

CHAPTER II. RELATED WORK 8

Methods for Accelerating Transcription 8

Optical Character Recognition for Ancient Manuscripts 8

Crowdsourcing Transcriptions for Ancient Manuscripts 9

Methods for Accelerating Identification 12

Pairwise Sequence Alignment 12

Genetic Sequence Alignment 13

CHAPTER III. A COMPUTATIONAL PIPELINE FOR CROWDSOURCED TRANSCRIPTIONS OF ANCIENT GREEK PAPYRUS FRAGMENTS 15

Introduction 15

Methods 16

Preprocessing Stage 17

Aggregation of User Clicks into Consensus Clicks 19

Creation of Line Sequences from Consensus Clicks 20

Pipeline Evaluation 21

Evaluation of Consensus Letter Identifications from Users 22
Evaluation of Line Sequence Creation Component 23
Results 24
Consensus Letter Identification Evaluation Results 24
Line Sequence Creation Component Evaluation Results 25
Conclusion 26
Future Work 27

CHAPTER IV. IDENTIFICATION OF ANCIENT GREEK PAPYRUS FRAGMENTS USING GENETIC SEQUENCE ALIGNMENT ALGORITHMS 28
Introduction 28
Related Work 30
Applied Sequence Alignment 30
Accelerated Transcription and Identification 31
Methodology 33
Calculating a New Substitution Matrix 33
Evaluation 35
Simulated Fragments 35
Results 37
Conclusion 38
Future Work 39

CHAPTER V. CONCLUSION 40
Accelerated Transcriptions for Ancient Greek Papyrus Fragments 41
Accelerated Identification for Ancient Greek Papyrus Fragments 42
Future Work 43

BIBLIOGRAPHY 44

LIST OF TABLES

Table 1 – Average F_1 scores of consensus letter identifications in each handwriting style from both the kernel-based approach and stepwise aggregation approach. 24

Table 2 – Average fragment lengths, average edit distances, and error ratios for the 41 fragments used in the line sequence creation component evaluation. Error ratios are calculated by dividing the average edit distance by the average fragment length. 26

Table 3 – The Greek Letter Oriented Substitution Matrix (GLOSUM) is a substitution matrix that contains a score for every possible letter-pair alignment in the Greek alphabet. 35

Table 4 – Example of a simulated fragment with key variable modifications highlighted in red. 36

Table 5 – Percentages of identification for each subset of fragment sequences used in the evaluation. 37

LIST OF FIGURES

Figure 1 – A deteriorated papyrus fragment from the Oxyrhynchus collection. . 3

Figure 2 – An example of how the image extraction algorithm misidentified a portion of P.Oxy fragment 3076 as multiple fragments. 9

Figure 3 – An example of a vertical line in the margin of P.Oxy fragment 4611. 10

Figure 4 – The Ancient Lives transcription interface. 11

Figure 5 – An example alignment between “elegant” and “elephant”. 12

Figure 6 – The architecture of the computational pipeline used to create consensus transcriptions from user-clicks made through the Ancient Lives interface. 17

Figure 7 – An Ancient Lives fragment image (left). A digital consensus transcription for the same Ancient Lives fragment (right). 18

Figure 8 – A visualization of the processing performed by the pipeline using the same Ancient Lives fragment from Figure 7. 20

Figure 9 – A Precision-Recall graph that visualizes how the precision and recall of each consensus letter identification from both approaches align with respect to calculated F_1 scores in Table 1. 25

Figure 10 – A papyrus fragment from Oxyrhynchus. 29

Figure 11 – Example alignment of two genetic sequences by BLAST [3]. 30

Figure 12 – A section of the original papyrus and the crowdsourced transcription. 32

Figure 13 – Example match between a fragment (top line) and a portion of a known full-text manuscript (bottom line) in Greek-BLAST. 33

Figure 14 – Above is a plot representing the average identification performance of all simulated fragments in the vertical gap rate subset based on a vertical gap rate ranging from 0 to 10 and fragment lengths ranging from 10 characters to 150 characters. 38

CHAPTER I.

INTRODUCTION

Classical historians agree that the vast majority of written knowledge from the ancient Mediterranean (*i.e.*, Egypt, Greece, and Rome) vanished with the burning of the Ancient Library of Alexandria in 48 BC [25]. With the destruction of the library, much of the modern world's knowledge of ancient life has been established by studying ancient texts. Some written works of antiquity have survived today day through a process known as medieval transmission. In this process, works of antiquity were manually selected and copied by scribes until the Middle Ages, where copies of these works finally made their way to print, thus being extant today. However, the works that did not undergo medieval transmission have since been lost.

Today, professionals continue to research the ancient world, specifically Graeco-Roman Egypt, by studying collections of papyrus fragments that were discovered and excavated from archaeological dig sites in the Mediterranean region. Over the last century, only a handful of ancient papyri collections have been found throughout the world. One of the most influential and richest collections of discovered papyrus to date is the Oxyrhynchus papyri[1], a vast trove of over 500,000 Greek papyrus fragments discovered in the ancient Egyptian city of Oxyrhynchus by British classicists Bernard Grenfell and Arthur Hunt [8]. After transporting the collection to the University of Oxford, these scholars and their successors began the development of formal methods and techniques for studying and extracting information from the papyrus fragments. Their efforts and work with the Oxyrhynchus papyri, alongside those of other specialists of the 19th and 20th centuries, would ultimately form the basis for the modern discipline of papyrology, the practice of studying ancient texts written on papyrus in order to better understand the culture, literature, and lifestyle of the ancient world, especially Graeco-Roman Egypt.

[1] http://www.papyrology.ox.ac.uk/POxy/

Papyrology

In working with a papyrus collection, a papyrologist, a practitioner of papyrology, first aims to transcribe each papyrus fragment. A papyrus fragment has been successfully transcribed once each available letter on the fragment has been confidently identified and matched to a known alphabetic letter. In the case of papyrus fragments, letters are usually restricted to those of the Greek, Latin, Coptic, Arabic, Aramaic, and Egyptian alphabets.

Despite being uniquely different from one another by content, modern papyrus collections are composed of fragments with partially-preserved information due to natural degradation of the papyrus material or exposure to extreme natural environments. In most cases, these fragments have a number of holes or gaps throughout the papyrus (see Figure 1). Additionally, the writing on a fragment may have been written illegibly, or the ink might have become less noticeable over time. This might cause confusion in distinguishing letters with similar shapes (*i.e.*, δ and λ) or recognizing the presence of letters at all. Collectively, these factors make the task of transcription much more difficult, even for professionals.

After a fragment has been transcribed, a papyrologist attempts to identify the anonymous papyrus fragment. The identification of an anonymous papyrus fragment has been completed when a papyrologist can confidently state that the content of the fragment belongs to a known manuscript. The first step in identifying a papyrus fragment is classifying the fragment as either literary or documentary. Literary papyrus fragments are written copies, often identical, of established literary works from the ancient world (*e.g.*, *Odyssey* by Homer). Conversely, documentary papyrus fragments are non-literary fragments whose contents contain non-literary information (*e.g.*, a bill of sale) and cannot be identified to a known author for this reason. A papyrologist can classify a fragment as documentary or literary based on the fragment's content, marginal features, and handwriting style. Generally, cursive handwriting signifies a documentary papyrus fragment, and non-cursive handwriting is representative of a literary papyrus fragment. In some cases, a fragment might be written in semi-cursive

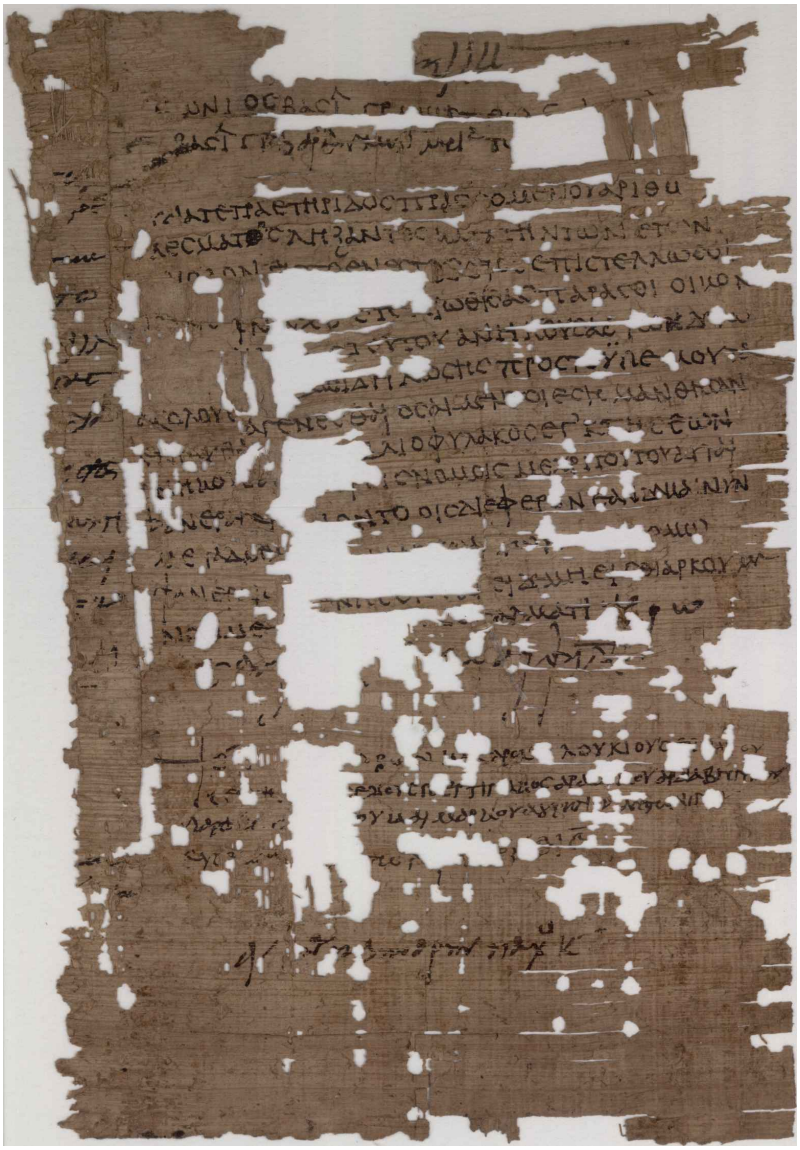

Figure 1: A deteriorated papyrus fragment from the Oxyrhynchus collection.

which would require additional attention to classify as literary or documentary.

Once a fragment has been classified as literary, the identification process is carried out by searching known literary manuscripts that contain the papyrus fragment's transcription. The success of the identification process is dependent on the correctness of the fragment's transcription, the extremity of the fragment's deterioration, and the papryologist's ability to quickly recognize known literary works. Like transcription, the task of identification is made more difficult by the deteriorated state of each fragment.

Thesis Statement

Due to the deteriorated nature of most papyrus fragments, the tasks of papyrological transcription and identification are extremely slow and tedious. For example, since its excavation in 1896, only about 10% of the more than 500,000 Oxyrhynchus papyri have been given preliminary identifications, and an even smaller percentage have been published. Two new computational approaches have been developed to accelerate the papyrological process.

To allow transcriptional crowdsourcing applications to accommodate the data collection needs of ancient texts, we developed a new computational pipeline to calculate accurate papyrus transcriptions from a large number (*e.g.*, millions) of crowdsourced letter identifications where each identification has an associated two-dimensional plane coordinate [39]. The pipeline consists of two components. The first component contains two approaches, the stepwise approach and the kernel-based approach, for aggregating and clustering identifications into consensus identifications for each fragment. An evaluation of each implementation using professionally verified classifications suggested that the stepwise approach is slightly more effective for calculating consensus letter identifications for non-cursive, semi-cursive, and cursive fragments. In addition, the stepwise approach seems more practical than the kernel-based approach as the stepwise approach executes approximately 576 times faster than the kernel-based approach. Using the consensus identifications from the first com-

ponent, the second component creates strings that represent digital transcriptions of the respective fragment. An evaluation of this component suggests that the component's underlying approach for calculating lines was effective for papyrus fragments with parallel lines. For 30 randomly selected fragments with parallel lines, the edit distance between the calculated transcription and the known transcription was on average 15.6% of the relative fragment's length. No known prior approach currently exists for automatically aggregating individual letter transcriptions made in two-dimensional space.

To accommodate the task of identification, a novel methodology is presented for leveraging genetic sequence alignment algorithms to the task of identifying an anonymous papyrus fragment to a known literary manuscript [38]. A demonstration of this methodology has culminated in Greek-BLAST, a variant of the BLAST algorithm, that can be used to identify deteriorated papyrus fragments to known literary manuscripts. In an evaluation using simulated fragments, Greek-BLAST identified 88.4% of the fragment queries as the highest-scoring identification. Nearly all cases of misidentification can be attributed to fragment length as simulated fragments with a length less than 10 characters were consistently unidentifiable. Once the length of the fragment had been extended to 20 characters, the fragment sequence was identifiable despite being modified severely by multiple key variables. Greek-BLAST accelerates the process of papyrus identification from multiple days, weeks, or months to a few seconds for a single fragment. This is the only known work that aims to explicitly accelerate the task of identification for deteriorated papyrus fragments.

In combination with one another, these two approaches accelerate the papyrological process well beyond the tedious manual process that currently exists and sets a new standard for the role of computing not only in papyrology, but across the arts and humanities in general.

Thesis Outline

This thesis is organized as the following: Chapter 2 covers previous work, detailing the current state-of-the-art techniques for accelerating the tasks of papyrus transcription and identification. Chapters 3 and 4 are conference papers that are detailed below. Finally, Chapter 5 discusses the conclusions of this thesis and suggests the direction for future work.

Chapter 3 is the conference paper, *A Computational Pipeline for Crowdsourced Transcriptions of Ancient Greek Papyrus Fragments*. This paper was presented at the 2014 IEEE International Conference on Big Data and was published by IEEE in the conference's proceedings. This paper compares two implementations of a computational pipeline for calculating crowdsourced transcriptions of ancient texts from a large number of crowdsourced letter identifications given by volunteers with varying levels of expertise. Each implementation is evaluated on the accuracy of calculated consensus letter identifications and calculated consensus transcriptions as compared to their professionally-verified counterparts. The results of the evaluation suggest that one implementation, the stepwise approach, is both a fast and accurate solution for calculating crowdsourced transcriptions from two-dimensional plane coordinates. Additionally, the approach is not bound by any specific alphabetic language and can be used to calculate transcriptions from any dataset of crowdsourced letter identifications with associated two-dimensional plane coordinates.

Chapter 4 is the conference paper, *Identification of Ancient Greek Papyrus Fragments Using Genetic Sequence Alignment Algorithms*. This paper was presented at the 10th IEEE International Conference on e-Science and published by IEEE in the conference's proceedings. This paper introduces a methodology for leveraging genetic sequence alignment algorithms to the task of identification for deteriorated papyrus fragments. The new methodology is used to calculate a new substitution matrix, the GLOSUM matrix, using misidentification statistics from Ancient Lives and letter-frequency statistics calculated by studying the Perseus database of ancient Greek manuscripts [34]. To test the effectiveness of

this approach, we have developed Greek-BLAST, a modified version of the popular genetic sequence alignment tool BLAST that includes the GLOSUM matrix. Greek-BLAST's ability to identify papyrus fragments was measured using simulated papyrus fragments that were modified based on multiple key variables that emulate deterioration. The successful identification of each simulated papyrus fragment was used to study which key variable was most detrimental to the identification process and to understand the practicality of the methodology. The results of the evaluation suggest that fragments containing vertical gaps are more difficult to identify than fragments subjected to any other key variable. Additionally, the results indicate the approach was successful for an average of 88.4% of the 14,100 simulated fragments. The success of the methodology is causal for future investigation as no other computational method currently exists for accelerating the task of literary papyrus identification.

CHAPTER II.

RELATED WORK

This section details the current state-of-the-art methods and techniques for hastening the processing of papyrus transcription and identification.

Methods for Accelerating Transcription

Optical Character Recognition for Ancient Manuscripts

With the increasing number of document digitization efforts, optical character recognition (OCR) algorithms have become a widely used technology for extracting textual information from documents. Today, OCR methods and techniques are used for a large variety of applications throughout society, such as commercial store check-out systems, office scanners, and the United States Postal Service [27]. They have also been extensively applied to historical texts and manuscripts [7, 36]. However, only one study to date suggests that an OCR-based approach for extracting text from deteriorated manuscripts is feasible [15]. Like most automated machine-driven techniques, the approach requires consistency in the image dataset's quality and color. For large collections of digitized texts, this might not be the case as the timeframe in which the digitization took place could have spanned multiple years or decades resulting in a image set of varying quality and possibly color as well.

In the case of the Oxyrhynchus papyri, fragments were photographed in groups of up to twenty fragments at a time depending on the size of each fragment. The image containing multiple fragments is called a sheet image. Several years ago, an algorithm was applied to each sheet image to extract and save each fragment in a separate image file. While the algorithm was effective at segmenting and extracting unique fragments, some of the fragment images lost important information due to the algorithm's inability to distinguish between lighter areas of the papyrus and the white background of the images that separates fragments (see Figure 2). Additionally, the algorithm unintentionally destroyed the original

(A) Before (B) After

Figure 2: An example of how the image extraction algorithm misidentified a portion of P.Oxy fragment 3076 as multiple fragments.

sheet images, making the process of extraction unrepeatable. A number of pre-processing approaches have been developed to account for document inconsistencies in images (*i.e.*, orientation) [12, 14], but due to the unintentional and irrecoverable deterioration from the extraction algorithm, even state-of-the-art OCR approaches are incapable of distinguishing letters in a large majority of the Oxyrhynchus papyri.

Crowdsourcing Transcriptions for Ancient Manuscripts

Since the debut of widespread personal internet access, crowdsourcing has emerged as an effective solution for leveraging human intelligence to solve computational problems that are beyond the scope of existing artificial intelligence algorithms. For ancient texts and manuscripts, crowdsourcing overcomes the complications that come with an OCR-based approach by enlisting humans with the task of recognizing characters. Additionally, some studies have shown that crowdsourcing can be more cost-effective than the development of automated approaches [20, 21].

Transcription has become an extremely common task in crowdsourcing applications. Some of the most successful transcription projects of today include Old Weather[1], Operation War Diary[2], and the Smithsonian's Transcription Center[3], an online platform for crowdsourc-

[1] http://www.oldweather.org/
[2] http://www.operationwardiary.org/
[3] https://transcription.si.edu/

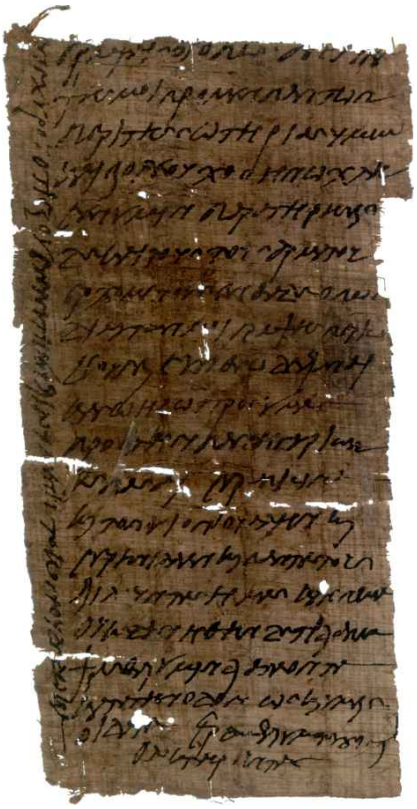

Figure 3: An example of a vertical line in the margin of P.Oxy fragment 4611.

ing the digitization of manuscript collections held by the Smithsonian. Each of these projects allow users to transcribe the document's contents in plain-text input fields where a field exists for each line of text in the document. Once enough transcriptional information has been collected, all user transcriptions are aggregated into a consensus transcription for each subject based on some rule for determining consensus (*e.g.*, majority vote). The resulting consensus transcriptions are representative of the final transcriptions for each document.

There are a number of difficulties associated with crowdsourcing historical texts, especially those from antiquity. First, plain-text input fields are often incapable of capturing transcriptions of secondary document content, such as marginal content or short notes written on the side of a page (see Figure 3). In cases where plain-text fields are used to capture this type of information, users might contribute dramatically different variations of tran-

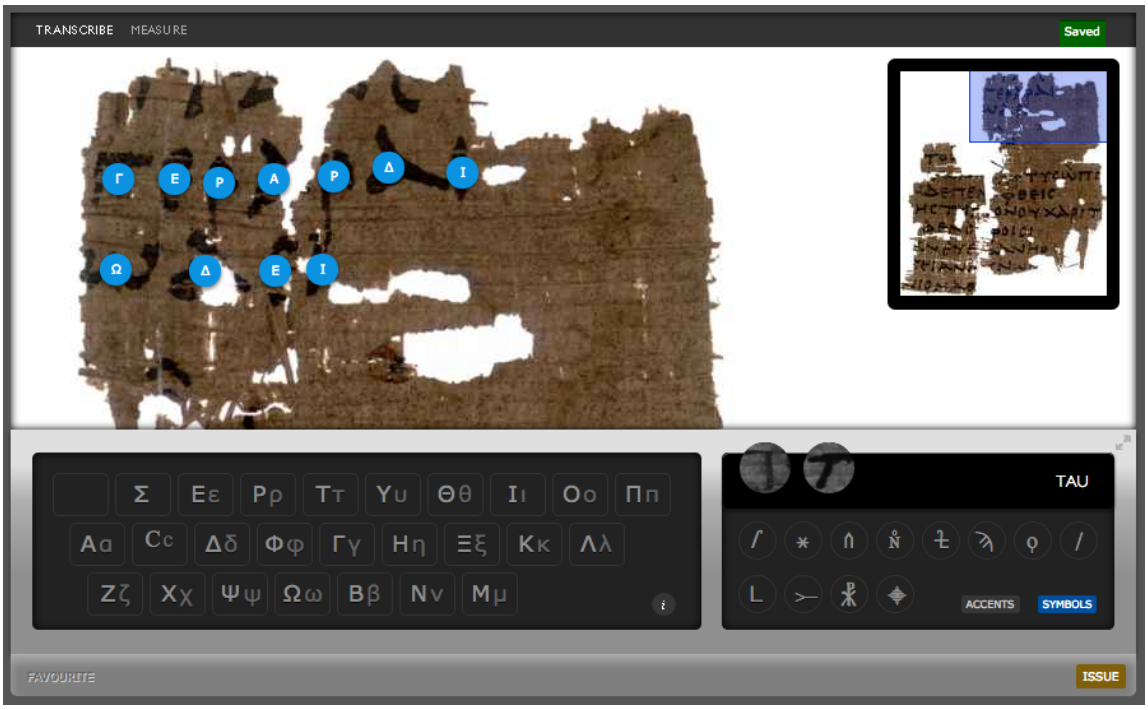

Figure 4: The Ancient Lives transcription interface.

scription, which can complicate the task of determining consensus. Furthermore, marginal content can be extremely relevant to understanding the context of the text. For example, in papyrology marginal content can potentially change the interpretation of an entire line of text. Second, over the past 25 years, the Unicode standard has been progressively adding support for ancient languages. Even with the added support for these languages, many modern keyboards are incapable of typing the necessary characters or letters to transcribe ancient texts and manuscripts.

More recent crowdsourcing applications have developed alternative methods of collecting transcriptions of ancient texts. For example, in the Ancient Lives transcription interface, users identify the presence of letters by clicking on the papyrus image and assigning a letter to that location using an on-screen keyboard (see Figure 4). By allowing users to supply transcriptions in a spatial dimension, all content of a document, including the marginal

content and non-traditional lines of text (*i.e.*, vertical lines of text), can be collected without problem. Additionally, by recording time-stamps with each classification, this method of collection allows researchers to analyze the order of each user classification and better understand how users perform the task. Despite overcoming complications with traditional data collection methods for historical texts, no known method or technique presently exists that can calculate transcriptions using letter identifications collected through this interface.

Methods for Accelerating Identification

Pairwise Sequence Alignment

Despite the lack of prior work aiming to specifically accelerate the process, the task of identifying deteriorated papyrus fragments bears great resemblance to the fundamental problem of finding regions of similarity between sequences of text. In computer science, pairwise sequence alignment is the formal approach of aligning two sequences in order to find regions of similarity (see Figure 5). In sequence alignment, the similarity of two sequences is defined as the score of their alignment. An alignment's score is determined by a scoring schema that assigns a score for each matching or mismatching letter-pair in an alignment. Scoring schemas for sequence alignment are a topic that has been studied rigorously, particularly in computational biology [2, 13, 18, 19]. In addition to matches and mismatches, an alignment of two sequences can have gaps, which represents an arbitrary number of consecutive insertions or deletions in either sequence. The presence of a gap in an alignment is usually penalized heavily in comparison to a mis-match as it represents one or more letters that do not contribute to the alignment.

```
e l e g - a n t
| | | :   | | |
e l e p h a n t
```

Figure 5: An example alignment between "elegant" and "elephant".

The two fundamental algorithms for pairwise sequence alignment are the Needleman-Wunsch algorithm [30] and the Smith-Waterman algorithm [35]. The Needleman-Wunsch algorithm is a well-known dynamic programming algorithm for calculating the optimal alignment. For every possible combination of gaps, the algorithm computes the maximum possible score for inserting a gap in the first sequence, inserting a gap in the second sequence, and finding a match in the alignment. The Smith-Waterman algorithm is a variant of the Needleman-Wusnch algorithm that allows segments of arbitrary length from each sequence to be aligned without being penalized for the unaligned regions of each sequence. In contrast with the stringency of the Needleman-Wunsch algorithm, the Smith-Waterman algorithm provides more leniency in the requirements of sequences having similar length and structure. Both the Needleman-Wunsch and Smith-Waterman algorithms are much faster than the traditional brute-force approach as the dynamic programming implementation dramatically reduces the number of required steps for computing alignments.

While both the Needleman-Wunsch algorithm and Smith-Waterman algorithm ensure the discovery of the optimal alignment at a rate faster than a standard brute-force approach, neither algorithm is practical for finding alignments between a query sequence and a database containing a large number of sequences. In order to accelerate the alignment process, heuristics (*e.g.*, *k*-tuple methods), were incorporated into new sequence alignment algorithms, such as FASTA [24] and BLAST[3]. The addition of the heuristic allows these algorithms to approximate the Smith-Waterman algorithm and operate at a dramatically shortened execution time, which is critical in working with large databases. However, as with the presence of any heuristic, this approach promotes the potential identification of nonoptimal alignments as the approximation reduces performance in accuracy [31].

Genetic Sequence Alignment

Despite their extensive use in a variety of fields [4, 6, 22], the large majority of sequence alignment research has come from studies in computational biology and bioinformatics [28].

In computationally biology, genes are represented as strings of letters, or genetic sequences, where each letter represents a specific amino acid or nucleotide residue. Due to the nature of genetic mutation, amino acids and nucleotides can transform, or mutate, to other amino acids or nucleotides. In the context of genetic sequences, mutations are characterized by the insertion, deletion, or substitution of amino acid or nucleotide letters in a genetic sequence. Many genetic sequence alignment algorithms allow users to specify custom gap penalties in order to overcome problems in alignment accuracy due to the unpredictability of gaps. In practice, a computational biologist can use a sequence alignment algorithm to attempt to identify an anonymous genetic sequence to known genetic sequences. As sequence similarity is indicative of shared evolutionary traits in biology, the identification of the anonymous sequence allows the computational biologist to better understand the origin of the gene and determine whether the gene has been identified before.

From the perspective of text, there are a number of similarities between genetic sequences and transcriptions of papyrus fragments. First, a biological mutation is equivalent to an error in the papyrus transcription in terms of primitive string operations of insertion, deletion, and substitution. With large numbers of torn edges, even professionals have difficulty distinguishing the presence of a letter. Furthermore, the accuracy in transcription is heavily dictated by legibility in handwriting and visibility of the papyrus' ink. Collectively, these issues amount to misidentified and missing letters. Second, the concept of a biological gap is analogous to physical tears and letters that cannot be transcribed in the papyrus fragment. Like in any physical text, tears and other harmful forms of deterioration can unpredictably happen horizontally, vertically, diagonally, or multidirectionally. Additionally, tears can span the entire length of a document, which can be extremely inimical to the fragment's already limited information. Despite the similarities between application areas, there has been no known previous effort to understand how genetic sequence alignment algorithms can be leveraged to the task of deteriorated papyrus identification.

CHAPTER III.

A COMPUTATIONAL PIPELINE FOR CROWDSOURCED TRANSCRIPTIONS OF ANCIENT GREEK PAPYRUS FRAGMENTS

Introduction

Over a century ago, two excavators, B.P. Grenfell and A.S. Hunt, of the University of Oxford, uncovered a vast trove of papryi, numbering over 500,000 fragments, from the city of Oxyrhynchus [8]. After transporting the collection back to the university, the field of papyrology emerged. Grenfell and Hunt began transcribing and editing the papyrus fragments, and to this day only a fraction of this vast trove have been published. Transcribing the collection has not been a simple task as each fragment suffers a unique level of deterioration with varying subsections of missing papyrus or illegible handwritten text. Due to the meticulous process of transcribing fragments with limited information, the rate of manual transcription for fragments is extremely slow. In order to quicken the process of transcription, the University of Oxford enlisted volunteers by establishing Ancient Lives[1], a web-based interface for identifying letters on digital images of papyrus. Users can log onto the Ancient Lives site and help transcribe ancient papyrus fragments by clicking on a location in the image and designating the presence of a specific letter. Each letter identification and its associated characteristics (*i.e.*, x,y coordinates) are stored in a database of user identifications. To date, over 7 million letter identifications have been been recorded internationally via the Ancient Lives interface.

In other projects that confront the task of transcribing historical documents, such as Transcribe Bentham [29], user transcriptions are given in plain-text with supplemental XML tags. However, for Ancient Lives, user transcriptions cannot be given in plain-text due to the absence of certain Ancient Greek characters and accents on the modern keyboard. In substitution of the physical keyboard, users use an on-screen keyboard that has the characters

[1] https://ancientlives.org

and accents necessary to transcribe any ancient Greek papyrus fragment. By capturing letter identifications through click data and the on-screen keyboard instead of plain-text, the interpretation of the letter identification data has become a nontrivial task. In order to more easily interpret the large amount of letter identification data, we present a new computational pipeline for automating the process of converting the crowdsourced letter identifications into digital consensus transcriptions of papyrus fragments. The Methods subsection details the design of each pipeline component in depth. The Evaluation subsection presents an assessment of each pipeline component performed by comparing the consensus letter identifications and consensus line sequences for a set of fragments in Ancient Lives to the fragment transcriptions and sequences for the same fragments as they appear in published Oxyrhynchus (P. Oxy.) volumes. The Results subsection provides an interpretation of the results found in each evaluation. The Conclusion subsection reiterates on the value of the pipeline and ends with discussion on future work.

Using the digital consensus transcriptions from the pipeline, both professional papyrologists and papyrology students will be able to more quickly and easily begin the papyrological process. Despite being designed specifically for the domain of papyrus, additional classification projects that are tasked with forming consensus letter identifications or consensus line sequences from data-click coordinates can make use of the pipeline architecture.

Methods

The computational pipeline can be separated into two stages (see Figure 6):

Stage 1: Aggregating User Clicks into Consensus Clicks.

Stage 2: Creating Line Sequences from Consensus Clicks.

The pipeline begins with a collection of plain-text files where each file contains click-data information for a specific fragment. The pipeline processing requires no human intervention after the input has been given. Once processing has finished, the pipeline will yield two files for each papyrus fragment. The first file contains the relative fragment's consensus letter

identifications with consensus x,y coordinates. The second file, which is the final output of the pipeline, contains the relative fragment's consensus line sequence that closely resembles the original papyrus fragment (see Figure 7). The consensus letter identification components are written in both Matlab and Python and the line sequence creation component is written only in Python. All supplemental visualization in Python is performed with version 2.4.9 of the OpenCV image processing and computer vision package[2].

Preprocessing Stage

The Ancient Lives interface is directly linked to a MySQL relational database that houses all transcription information. For every click a user makes on a digital papyrus image, the database will store the unique user-id of the "citizen", the user's relative click location for the letter (*i.e.*, x,y coordinates), and the citizen's letter choice in unicode. From the database, we retrieve all click information and categorize user clicks into separate files for each fragment. Separating and organizing the click data by fragment allows us to more easily analyze click information on the basis of individual fragments. More specifically, the procedure encourages the detection of strong dissimilarities between individual user clicks and the consensus clicks for a given fragment (*i.e.*, an accidental click on the image).

In some cases, a digital papyrus image is incorrectly oriented in the Ancient Lives

[2]https://opencv.org

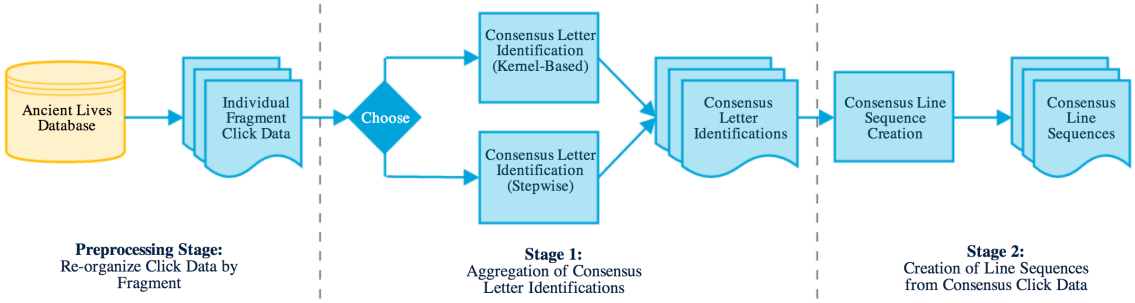

Figure 6: The architecture of the computational pipeline used to create consensus transcriptions from user-clicks made through the Ancient Lives interface.

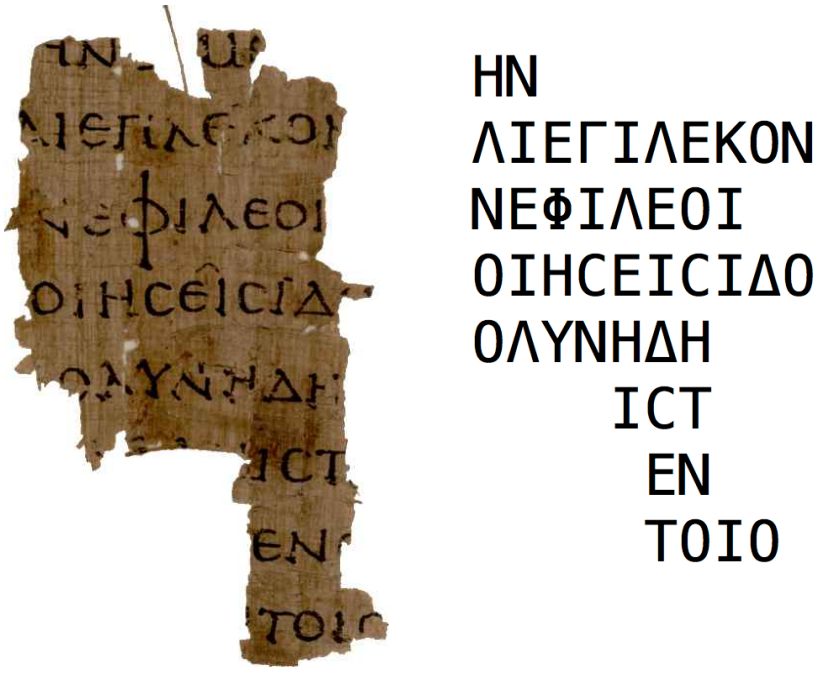

Figure 7: An Ancient Lives fragment image (left). A digital consensus transcription for the same Ancient Lives fragment (right).

database or a user might transcribe a fragment sideways while maintaining an accurate transcription. Subsequently, the click-data coordinates that are relative to the image are also incorrectly oriented. Where applicable, processing rotations in click-data is a necessary step in order to ensure line sequences are correctly formed. From the Ancient Lives MySQL database, we also query relevant rotation information (*i.e.*, the rotation degree used by users to transcribe) for each fragment. Using the rotation information retrieved from the database, we determine the correct rotation degree of each fragment by identifying the most frequently used orientation among all users during the transcription process. Afterwards, all click-data undergoes a rotation filter that adjusts x,y coordinates based on the identified orientation degree.

Aggregation of User Clicks into Consensus Clicks

Two unique approaches were developed for the task of aggregating all user clicks for a given fragment to form consensus identifications for letters. We refer to the first approach as the kernel-based approach. This approach was written in Matlab and leverages kernel density estimation, a mathematical approach for inferring the likelihood that a variable will take on a given value, to identify consensus clicks and letters [17]. The algorithm begins by distributing all user click data into a number of bins based on the click's x,y coordinates. The number of bins is determined by multiplying a user-specified kernel width by 2. If no kernel width is specified by the user, the kernel width is assigned a default value of 8. Within each unique bin, the algorithm will identify the highest kernel density peaks, which represent the presence of a consensus letter. Once peaks have been identified within each bin, a filtering function is imposed to prevent duplicates and eliminate suspected false consensus letters. The x,y coordinates of the remaining kernel density peaks are clustered and used to determine the location of consensus letters.

Due to the nature of calculating the kernel density estimation for millions of user clicks, the kernel-based approach requires a large amount of computational overhead and takes multiple days to process user click data to yield consensus letter identifications. In order to hasten the processing time, a second approach, referred to in this paper as the stepwise aggregation approach, was developed in Python. This approach relies on a recently established concept that citizen scientists who complete more classification tasks have an elevated level of knowledge and reason in classifying data correctly than those who complete fewer classification tasks [32]. Based on the concept that expertise can be represented by experience or frequency of activity, the algorithm will first identify the user that has made the highest number of clicks on the fragment and use their clicks as seed locations for potential consensus letter identifications. Depending on an unprocessed click's proximity to pre-existing seed locations, the remaining user clicks are either merged with a preexisting

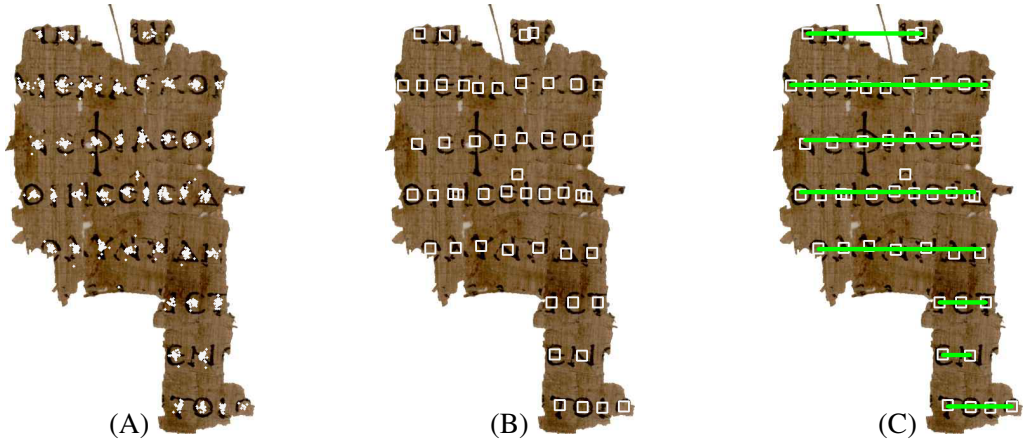

Figure 8: A visualization of the processing performed by the pipeline using the same Ancient Lives fragment from Figure 7. A) White dots represent the 1,591 clicks of all users. B) White squares represent calculated consensus clicks from the aggregated click data of all users for the fragment using the stepwise aggregation approach . C) Green lines represent the regression line for the nearby consensus clicks.

seed location or used to establish a new consensus letter location and added as a seed location. Once all user clicks have been processed, a centroid, or center point, of each agglomeration of clicks is identified and recorded as a consensus letter (see Figure 8B).

Creation of Line Sequences from Consensus Clicks

Using the consensus letter identifications from the previous stage, the line sequence creation component will attempt to form line sequences that closely resemble the text presented in the digital image of the papyrus fragment. The input of line sequence creation component is a text file containing a list of x,y coordinates with associated Greek characters. The output of the line sequence creation component is a text file containing a line sequence, or string, composed of all of the characters that appear in the input file.

The algorithm begins by sorting all clicks into a list based on the y coordinate. Beginning at a y-coordinate of 0 and ending at a y-coordinate equal to the height of the relative fragment's digital image, the algorithm searches the sorted y-coordinates and identifies the

presence of lines based on gaps of vertical space between neighboring y-coordinates (see Figure 8C). When a line is identified, the y-coordinate is added to a list of line regions. After each click has been grouped into a line region based on its relative x,y coordinate, the best fit line, or regression line, for each line region is calculated. In addition to the equation, the average space between neighboring line regions is calculated. Using the equation of the best fit line as a reference, a second pass of all y-coordinates is made in order to ensure that each letter was categorized in the correct line region. If an x,y coordinate is not within half of the calculated average space between neighboring line regions from the relative line's median y-coordinate, the coordinate is categorized into another line region. After the second pass of y-coordinates has finished, each line region is sorted by the x-coordinate in order to ensure characters appear in the same order they appear on the papyrus. After the regions have been sorted by x-coordinate, the regions are concatenated into a single string, which represents the line sequence for the fragment.

There are two types of styles of line that are presented in papyrus. The first style is parallel where lines are written in straight, distinct, and predictable lines and are equidistant from neighboring lines. The second style is curvilinear where lines are written in the shape of an arc and are unpredictable in direction. For most papyrus fragments with curvilinear lines, identifying a consistent amount of vertical space between line regions is nontrivial. As a result, the described approach could produce duplicate lines by identifying multiple line regions from a single curvilinear line due to incorrect measurements of vertical space between line regions. In order to filter duplicate line regions, a final post-processing stage will remove a line region if it shares 70% or more identity with its neighboring line region.

Pipeline Evaluation

In this evaluation, we examine the efficacy of Stage 1 and Stage 2 separately. In addition to fragments that have yet to be transcribed, a group of published fragments with known transcriptions were deposited into Ancient Lives in order to make assessing the effectiveness

of each pipeline component possible. Given that each is supplied with the same set of click data, both the kernel-based approach and the stepwise aggregation approach are scrutinized on the ability to correctly classify a letter by comparing consensus click data to the click data of published fragments given by a professional papyrologist. Similarly, the performance of the line sequence creation component is evaluated based on the similarity between line sequences produced by the component and digital fragment transcriptions as they appear in a published P. Oxy. volume.

Evaluation of Consensus Letter Identifications from Users

We randomly selected 54 published fragments from the Oxyrhynchus collection to be used as evaluation criteria for measuring the accuracy of consensus letter identifications from both the kernel-based approach and stepwise aggregation approach in comparison to known letter identifications that appear in P. Oxy. volumes. In this assessment, the default kernel width value of 8 is used in the kernel-based approach. In order to evaluate on the basis of user clicks made by citizen scientists, clicks made by professional papyrologists have been removed from the click data set used to make consensus letter identifications. Each fragment was categorized based on handwriting style and legibility into one of the following groups: non-cursive, semi-cursive, and cursive. We utilize three established metrics, precision, recall, and F_1 score [37], for determining the classification performance of each approach. Precision is calculated by dividing the number of correct letter identifications by the number of total letter identifications in the consensus transcription. Recall is calculated by dividing the number of correct letter identifications by the total number of letter identifications in the relative fragment's P. Oxy. transcription. Both precision and recall are combined into a composite metric, the F_1 score. The equation to calculate the F_1 score for an individual fragment is:

$$F_1 = 2 \times \frac{P \times R}{P+R}$$

where P and R are respectively the precision and recall of the fragment's click data. A F_1 score of 0.0 can be interpreted as an approach correctly classifying next to none of the letters while a F_1 score of 1.0 can be interpreted as an approach correctly classifying most or all of the letters.

Evaluation of Line Sequence Creation Component

In this evaluation, the accuracy of the line sequence creation component is examined by supplying the component with professionally curated click-data, where each click represents a correct letter at the correct relative x,y coordinate. From the same set of fragments used in the previous evaluation, a subset of 41 fragments were selected to be used as evaluation criteria to examine the sequence similarity of consensus line sequences produced through the line sequence creation component and the published digital transcription of the same fragment. All 41 fragments were chosen on the basis that a digital transcription exists for the fragment. For the remaining 13 fragments in the set of fragments used in the previous evaluation, a digital transcription does not exist. Each fragment with a digital transcription was categorized as having either parallel lines or curvilinear lines. Of the 41 fragments, 30 were categorized as having parallel lines and 11 were categorized as having curvilinear lines. The edit distance, or Levenshtein distance [23], is employed as a metric to measure the similarity between the transcription produced through the line sequence component and the transcription that appears in a P. Oxy. volume. An edit distance of 0 represents complete identity between two strings (*i.e.*, an exact match). For an edit distance that is greater than 0, at least one insertion, deletion, or substitution was required for one sequence to resolve to the other.

Results

Consensus Letter Identification Evaluation Results

The results of the evaluation for Stage 1 suggest the stepwise aggregation approach produces a higher level of accuracy for correctly determining consensus letter identifications than the kernel-based approach, especially for fragments with cursive handwriting (see Table 1).

The small difference in performance between the kernel-based approach and the stepwise aggregation approach can be explained by how each method extrapolates on user clicks. In the stepwise aggregation approach, every user click is used to create the consensus click data set for a fragment. In the kernel-based approach, every user click is also used to create the consensus click data set, but in order to prevent duplicate letter identifications, the list of suspected consensus clicks undergoes a filtering process, which has the potential to remove true-positive letter identifications and decrease identification performance. The difference in accuracy of correctly classifying true-positives and true-negatives can be visualized with the precision and recall (see Figure 9). In addition to producing a higher level of accuracy, the stepwise aggregation approach has an accelerated execution time in comparison to the kernel-based approach. The total execution time for the stepwise aggregation approach is

Table 1: Average F_1 scores of consensus letter identifications in each handwriting style from both the kernel-based approach and stepwise aggregation approach.

Handwriting Style	**Average Kernel-Based F_1 Score**	**Average Stepwise Aggr. F_1 Score**
Non-cursive (34)	0.65	0.67
Semi-cursive (12)	0.64	0.68
Cursive (8)	0.61	0.69
Aggregated (54)	0.64	0.67

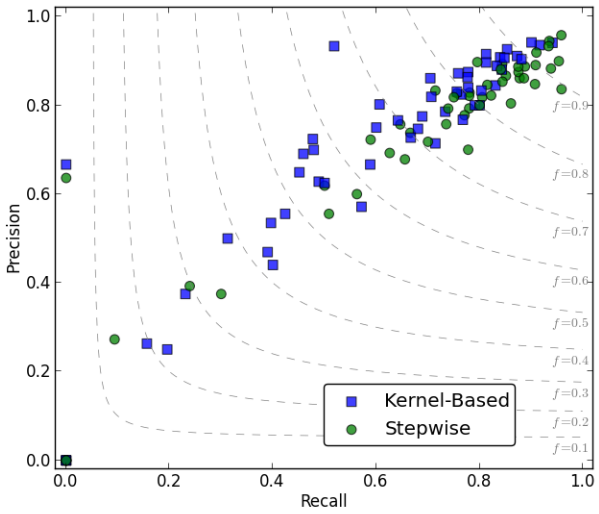

Figure 9: A Precision-Recall graph that visualizes how the precision and recall of each consensus letter identification from both approaches align with respect to calculated F_1 scores in Table 1.

currently about fifteen minutes while the execution time or the kernel-based approach spans a few days.

Line Sequence Creation Component Evaluation Results

The results of the evaluation for Stage 2 suggest that the line sequence creation component is effective at creating sequences correctly for most fragments with parallel lines (see Table 2). Of the 30 fragments categorized as having parallel lines, eleven sequences created through the line sequence component were exact matches to their digital transcription counterpart. Of the 11 fragments categorized as having curvilinear lines, only one sequence created through the line sequence was exactly matched to its digital transcription counterpart. There is a clear distinction between the component's effectiveness for fragments containing parallel lines and the component's effectiveness for fragments containing curvilinear lines. The difference in performance can be explained by the current approach's inability to consistently determine which letters belong to which line in fragments with curvilinear

Table 2: Average fragment lengths, average edit distances, and error ratios for the 41 fragments used in the line sequence creation component evaluation. Error ratios are calculated by dividing the average edit distance by the average fragment length.

Line Style	Average Fragment Length	Average Edit Distance	Error Ratio
Parallel (30)	43.7	6.8	15.6%
Curvilinear (11)	234.0	83.3	35.6%
Aggregated (41)	94.7	26.1	27.6%

lines.

Conclusion

In order to more easily interpret the large amount of letter classification data from over a million users, we presented a new computational pipeline for translating millions of user-clicks on digital images of Ancient Greek papyri to digital, consensus transcriptions that closely resemble the format of the original papyrus. Both professional papyrologists and student papyrologists can utilize the digital consensus transcriptions produced through the pipeline to more quickly examine, edit, and publish fragments with confidence. Despite being designed specifically for Ancient Greek papyrus fragments, classification projects that share the task of forming either consensus letter identifications or consensus lines of text from coordinate click-data can take advantage of the computational pipeline.

By engineering a pipeline for interpreting the millions of user-clicks from Ancient Lives, we are dramatically redefining how professional papyrologists and scholars interact with ancient papyri. Typically, a papyrologist could spend days, weeks, or months manually transcribing multiple papyrus fragments. Both the pipeline and the Ancient Lives project leverage the work of citizens in order to help the papyrologist more quickly transcribe and evaluate fragments. There are a number of computational systems that have already made use of the pipeline's output and share the goal of bringing ease to the transcription process.

Greek-BLAST [38], for example, is a variant of BLAST, a popular genetic sequence alignment tool, designed specifically for suggesting identifications for literary papyrus fragments. Consensus transcriptions of literary fragments made through Ancient Lives can be supplied to Greek-BLAST directly as input and quickly aligned with matches in Ancient Greek literary manuscript databases (*i.e.*, The Perseus Digital Library [34]). Additionally, collaborators at the University of Minnesota have developed a web-based tool for quickly curating the digital consensus transcriptions produced from the computational pipeline. Curated consensus transcriptions are stored in a database that will later be used for data mining purposes. Lastly, the consensus transcriptions made through Ancient Lives will be the basis for many fragments that will be further studied, edited, and published in *The Oxyrhynchus Papyri* volume series and Proteus, a new interactive, web-based platform that leverages advanced computational methods and techniques to both the study and analysis of ancient texts and the creation of next-generation digital editions.

Future Work

A key component of future work is improving the line sequencing stage of the pipeline. For fragments written in a curvilinear manner, forming lines is a nontrivial task. We will investigate measures to help identify the presence of curvilinear lines and how the accuracy of the existing approach for developing line sequences can be improved. A final re-design of the pipeline will take place after additional classification information (*i.e.*, methods for identifying line information or missing papyrus) is added to the Ancient Lives framework.

CHAPTER IV.

IDENTIFICATION OF ANCIENT GREEK PAPYRUS FRAGMENTS USING GENETIC SEQUENCE ALIGNMENT ALGORITHMS

Introduction

The history of the ancient world holds much information that, even today, has yet to be discovered and, perhaps more importantly, understood. Much of the modern work done to further understand the culture, history, and literature of the ancient world is performed by papyrologists who transcribe and identify fragments of both unknown and known ancient literature and other written works as preserved on ancient papyrus manuscripts. Papyrologists transcribe, identify, and edit papyrus fragments by manually recognizing characters and strings of text and matching them to known full-text manuscripts. Papyrologists can spend days, weeks, or even months on the transcription and interpretation of damaged ancient texts (see Figure 10). For example, 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 [8]. In severe cases, damaged texts may be missing a large number of words and as a result, the amount of information that a papyrologist can transcribe and interpret is extremely limited.

Past research has demonstrated that computational biology solutions can be usefully applied to data mining and machine learning problems [1]. Genes are often digitally represented by a sequence of continuous letters from a finite letter set, where each letter represents a specific nucleotide or amino acid. Relationships are inferred by finding multi-letter patterns, which can be separated by insertions, deletions, or gaps, shared between the anonymous sequence and a known sequence. These matches are scored based on how well they align with one another using a substitution matrix. A substitution matrix has a score for the likelihood of each pair of amino acids being aligned. If the alignment between any two sequences produces a score that meets a user-defined threshold, the relevant sequence pair is identified to the user. This process, or algorithm, is commonly referred to

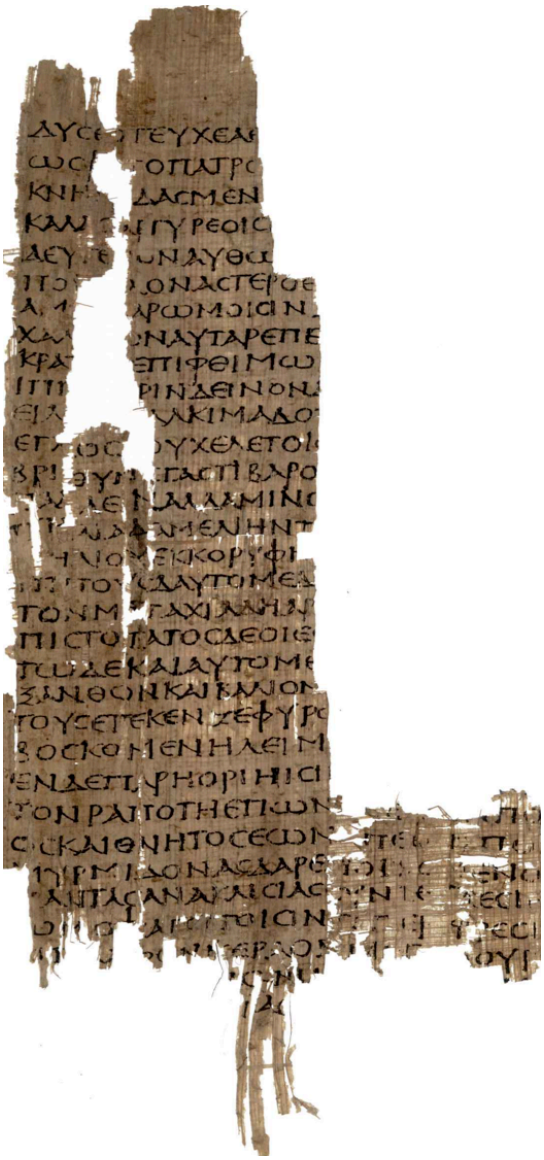

Figure 10: A papyrus fragment from Oxyrhynchus. There are obvious gaps in the image caused by degradation of the papyri. In some cases, letters have been partially lost due to gaps. Notably, most of the text on the right side is missing. The hand of the scribe is extremely clear in this literary fragment, but in many cases, the handwriting is more difficult to read.

as genetic sequence alignment (see Figure 11). In order to enable papyrologists with the ability to identify severely damage texts, we investigate the applicability of genetic sequence alignment algorithms as a method for fragmentary Ancient Greek text identification.

```
>ref|WP_006987218.1| oxidoreductase [Gillisia limnaea]
 gb|EHQ04326.1| short-chain dehydrogenase/reductase SDR [Gillisia limnaea DSM
15749]
Length=232

 Score = 98.6 bits (244),  Expect = 1e-23, Method: Compositional matrix adjust.
 Identities = 42/66 (64%), Positives = 57/66 (86%), Gaps = 0/66 (0%)
 Frame = -2

Query  199  VAPSITNTPLAQRLLSSSDKEEASAKRHPLHRVGKAKDIGSMAAFLLSDQSGWMTGQILG  20
            +APS+TNTPLA++LLS+ +K++   +RHPL RVG+AKDI +M  FLLS++S WMTGQ+LG
Sbjct  162  IAPSLTNTPLAEKLLSNDEKKKKMDERHPLKRVGEAKDIANMVVFLLSEKSSWMTGQVLG  221

Query  19   VDGGLS  2
            +DGGLS
Sbjct  222  MDGGLS  227
```

Figure 11: Example alignment of two genetic sequences by BLAST [3].

Providing professionals in the humanities with such a computational tool will dramatically accelerate the rate of papyri identification. Furthermore, this study outlines a new methodology for re-tailoring specialized sequence alignment tools, specifically those related to computational biology, to radically different textual domains. Instead of using a genetic sequence and database of known genetic sequences as input, our application will use the text from a Greek papyrus fragment and a database of complete, known Greek manuscripts. The texts on Greek papyri were written without word division and have little to no punctuation. Transcription data is essentially a string of Greek characters. As a result, the digital representation of recorded Ancient Greek texts is very similar to that of gene sequences. While genetic sequence alignment shares many similarities with Greek text fragment identification, a few key differences will be addressed to tailor the method to aid papyrologists.

Related Work

Applied Sequence Alignment

Sequence alignment algorithms are ubiquitous in text similarity search scenarios and have been used to provide interesting solutions to recent problems in natural language processing [26] and historical linguistics [33]. Past work has demonstrated that homology

search problems in computational biology can usefully take advantage of identical sequence alignment algorithms and techniques [5]. Unlike traditional sequence alignment algorithms, genetic sequence alignment algorithms have been tailored for the domain of amino acid and nucleic acid sequences. When using a genetic sequence alignment algorithm to identify and match sequences, it is common to find gaps in the alignment of a query sequence and a known genetic sequence. Gaps, symbolized by the '-' character in alignments, represent the insertion or deletion of a character, or characters, in one of the genetic sequences. The ability to account for new or missing information between aligned sequences is analogous to the problem of missing information in damaged or deteriorated papyrus fragments. Furthermore, modern genetic sequence alignment algorithms are highly parameterizable. For example, users may specify penalties, such as gap-penalties, to be used during the alignment scoring phase. These user-specified penalties allow sequence alignment algorithms to produce dramatically different results. By nature, genetic sequence alignment algorithms are tolerant to small inconsistencies in string similarity, which could be helpful in overcoming spelling mistakes and changes in inflection. We are not aware of any other research initiatives using genetic sequence alignment algorithms to identify papyrus fragments.

Accelerated Transcription and Identification

Two primary research efforts have aimed to hasten the tedious process of manual transcription. The first project is Oxford University's Ancient Lives project[1]. This project, like others in the Zooniverse[2], enlists volunteers to help process and analyze data. For the Ancient Lives project specifically, the input of thousands of volunteers are aggregated to aid in the deciphering the Greek letters from images of fragments (see Figure 12). While viewing images of papyrus fragments, these volunteers identify the Greek letters by clicking on a letter in the image. Since the volunteers are not trained scholars, they are asked to transcribe individual letters rather than to translate the manuscripts. By processing these

[1] http://www.ancientlives.org/

[2] http://www.zooniverse.org/

ΔΥCΕΟΓΕΥΧΕΑΕ
ΩCΙ ΤΟΠΑΤΡC
ΚΝΗ ΔΑCΜΕΝ
ΚΑΛΑΤΝΓΥΡΕΟΙC
ΔΕΥΓΕΝ ΝΑΥΘΩ
ΙΠΟ ΛΟΝΑCΤΕΡΟΕ
Α ΑΡΩ ΜΟΙCΙΝ
ΧΑΛ ΟΝΑΥΤΑΡΕΠΕ
ΚΡΑΝ ΕΠΙΦΘΙΜ Ω
ΙΠΙΙ ΡΙΝΔΕΙΝΟΝΑ
ΕΙΛ ΛΛΚΙΜΑΔΟ
ΕΓΛΟC ΟΥΧΕΛΕΤΟΙ
ΒΡΙΘΥΝ ΓΑCΠΒΑΡΟ

Figure 12: A section of the original papyrus and the crowdsourced transcription. The crowdsourced transcription is shown immediately next to a portion of the original image of the papyrus from Figure 10. The consensus transcript is used for document identification.

clicks, we obtain consensus textual versions of the fragments which can be utilized by our program. To date, this project has collected in excess of 7 million clicks from volunteers. For each papyrus fragment, the project collects the mouse-click and keyboard input of 5 to 20 users. The Ancient Lives project has greatly helped to accelerate the transcription process, but the task of identification still remains tedious.

The second project is the eAQUA project, which uses modern text mining techniques to extract structured knowledge from Ancient Greek texts [9]. The project's most recent contribution is a spell-checking system for Ancient Greek fragments. This new component is powered by natural language processing techniques that depend on semantics, syntax, and morphology. This system is capable of suggesting corrections for a single word (*e.g.*, a damaged or incorrectly transcribed word) in a fragment [10]. While valuable when a single word is damaged, its application is limited for large-scale, real-world use. For

```
Score = 68.4 bits (154),  Expect = 8e-13

Ancient Lives fragment: 131383

FRAGMENT ΠΑ?ΑΔ0?Ι?ΑΓΑθΗΚΑΙΠΑΝΔΩΡΗΜΑΤΕΛΕΙ0ΝΑΝΩθΕΝΕ?ΤΙΝΚΑΤΑΒΑ
TEXT     ΠΑΣΑΔ0ΣΙΣΑΓΑθΗΚΑΙΠΑΝΔΩΡΗΜΑΤΕΛΕΙ0ΝΑΝΩθΕΝΕΣΤΙ-ΚΑΤΑΒΑ
SIMILAR  ΠΑ ΑΔ0 Ι ΑΓΑθΗΚΑΙΠΑΝΔΩΡΗΜΑΤΕΛΕΙ0ΝΑΝΩθΕΝΕ ΤΙ ΚΑΤΑΒΑ

FRAGMENT ΙΝ0ΝΑΠ0Τ0ΥΠΑΤΡ0?ΤΩΝΦΩΤΩΝ
TEXT     ΙΝ0ΝΑΠ0Τ0ΥΠΑΤΡ0ΣΤΩΝΦΩΤΩΝ
SIMILAR  ΙΝ0ΝΑΠ0Τ0ΥΠΑΤΡ0 ΤΩΝΦΩΤΩΝ
```

Figure 13: Example match between a fragment (top line) and a portion of a known full-text manuscript (bottom line) in Greek-BLAST. Additionally, the statistics regarding the match are given below. Notice that the algorithm tolerates missing letters in the fragment.

damaged fragments, such as those found at Oxyrhynchus, content is presented with multiple incomplete or missing characters and words throughout the fragment.

Methodology

In order to leverage computational biology algorithms to Ancient Greek text fragment identification, we modified version 2.2.28 of the popular pairwise genetic sequence alignment algorithm, Basic Local Alignment Search Tool (BLAST) [3] (see Figure 13). This new BLAST variant has been appropriately named Greek-BLAST.

Calculating a New Substitution Matrix

In genetic sequence alignment algorithms, sequence alignments receive a final alignment score based on the scoring schema of letter-pairs defined in the substitution matrix. One of the most common families of substitution matrices in computational biology used by BLAST is the BLOSUM (BLOcks Substitution Matrix) matrix family [18]. The BLOSUM matrix family was empirically calculated by extracting ungapped sections of alignment from a database of observed genetic sequence alignments. Once the relative frequencies for each amino acid were calculated, a log-odds ratio was recorded for every possible amino acid substitution pair. The formula for constructing the BLOSUM matrix is:

$$S_{ij} = \frac{1}{\lambda} log\left(\frac{p_{ij}}{q_i q_j}\right)$$

where p_{ij} is the probability of two amino acids i and j replacing one another in any sequence and q_i is the background frequency for finding amino acid i in any sequence. S_{ij} is the index in the substitution matrix for the respective letters i and j and λ is a scaling factor. Multiple BLOSUM matrices were calculated based on different levels of similarity between the studied sequences. These matrices, the BLOSUM62 matrix in particular, have been validated as the best performing matrices for finding biologically relevant sequence alignments [19].

Using a similar log-odds methodology that was used to calculate the BLOSUM matrices, we introduce a new substitution matrix, the Greek Letter Oriented Substitution Matrix (GLOSUM). To calculate the target frequency (p_{ij}) for each letter pair, we studied the consensus letter identifications provided by the University of Oxford's Ancient Lives project. For each letter identification, we operated under the assumption that the consensus letter the correct letter. Any letter identification that did not match the consensus identification was treated as a misidentification and was used to create a matrix of misidentification percentages for each letter pair. We use these misidentification ratios as the target frequency for the log-odds formula. To calculate the background frequency (q_i), a proprietary database of 6,619 full-text Ancient Greek manuscripts, referred to here as the Training database, was studied to retrieve letter frequencies. The GLOSUM matrix was calculated from the log-odds ratio of the target frequency and background frequency for each letter pair. In order to amplify the positive scoring scheme for identical letter pair alignments and negative scoring scheme for non-identical letter pair alignments, the calculated matrix was summed with an identity substitution matrix where each index on the diagonal contained a score of 4 and all other indices on the matrix contained a score of -4 (see Table 3).

Table 3: The Greek Letter Oriented Substitution Matrix (GLOSUM) is a substitution matrix that contains a score for every possible letter-pair alignment in the Greek alphabet.

(α)	A	6																								
(ρ)	R	-4	7																							
(ν)	N	-4	-4	6																						
(δ)	D	-3	-4	-4	7																					
(ξ)	C	-4	-3	-4	-3	9																				
(θ)	Q	-4	-3	-4	-3	-3	8																			
(ε)	E	-5	-3	-4	-4	-2	-2	6																		
(γ)	G	-4	-3	-4	-4	-3	-3	-4	7																	
(η)	H	-5	-4	-4	-4	-3	-3	-4	-3	7																
(ι)	I	-4	-3	-4	-3	-2	-3	-4	-3	-3	6															
(λ)	L	-3	-4	-4	-2	-3	-3	-4	-3	-4	-4	7														
(κ)	K	-4	-4	-4	-4	-3	-4	-4	-4	-4	-4	-3	7													
(μ)	M	-4	-4	-4	-4	-3	-4	-4	-4	-3	-4	-3	-3	7												
(φ)	F	-4	-3	-4	-3	-2	-2	-4	-3	-4	-4	-4	-4	-4	8											
(π)	P	-4	-3	-4	-4	-3	-4	-4	-2	-3	-4	-4	-4	-3	-4	7										
(σ)	S	-5	-5	-6	-5	-2	-4	-4	-4	-5	-5	-5	-5	-5	-5	-5	6									
(τ)	T	-4	-4	-4	-4	-3	-4	-4	-2	-4	-4	-4	-4	-4	-4	-3	-4	6								
(O)	W	-4	-4	-4	-4	-3	-3	-5	-4	-4	-4	-4	-4	-3	-3	-4	-4	-4	7							
(ψ)	Y	-4	-3	-3	-3	-2	-2	-3	-2	-2	-3	-3	-3	-3	-1	-3	-2	-3	-2	10						
(υ)	V	-4	-4	-4	-4	-3	-3	-4	-2	-4	-4	-4	-4	-3	-3	-4	-4	-4	-4	-2	7					
(β)	B	-4	-3	-4	-3	-2	-2	-4	-4	-4	-4	-4	-3	-4	-3	-4	-4	-5	-4	-3	-4	8				
(ω)	J	-4	-3	-4	-3	-2	-3	-4	-3	-4	-4	-4	-4	-4	-3	-4	-3	-4	-3	-2	-3	-3	6			
(ζ)	Z	-4	-3	-3	-2	0	-3	-4	-3	-3	-3	-3	-3	-3	-3	-3	-2	-4	-3	-1	-3	-2	-4	9		
(χ)	X	-4	-3	-4	-3	-2	-3	-4	-3	-4	-4	-2	-3	-4	-3	-4	-4	-4	-4	-1	-3	-3	-4	-2	8	
		A (α)	R (ρ)	N (ν)	D (δ)	C (ξ)	Q (θ)	E (ε)	G (γ)	H (η)	I (ι)	L (λ)	K (κ)	M (μ)	F (φ)	P (π)	S (σ)	T (τ)	W (O)	Y (ψ)	V (υ)	B (β)	J (ω)	Z (ζ)	X (χ)	

Evaluation

In this evaluation, we utilized 14,100 fragment sequences as input to Greek-BLAST. The accuracy of each simulated fragment query was determined by whether or not the correct manuscript was presented as the highest scoring match in the list of relevant matches. If the manuscript to which the simulated fragment belonged to was not the highest scoring match, the query fragment was categorized as being unsuccessfully identified.

Simulated Fragments

To remove the potential for type II classification errors, we simulated fragments based on the Training database. From the database, 10 known manuscripts were randomly selected and used to create fragments with different levels of deterioration. Deterioration was emulated by extracting fragments, or substrings, of varying lengths from the manuscripts and modifying

Table 4: Example of a simulated fragment with key variable modifications highlighted in red.

Key Variable	Value	Resulting Fragment
Original		ωυδενααπτπω
Deletion Rate	0.2	ωυναπτπω (Removed δ,ε)
Ex. Char Rate	0.2	ωυδενααπατπωψ
Error Rate	0.2	ωψδενααωτπω
Vert. Gap Rate	2	ωυδεπτπω (Removed ν,α)

the fragment based on one of four key variables: error rates, vertical gaps, extra character rates, and character deletion rates. Error rate, which simulates an identification error made by an Ancient Lives volunteer, symbolizes the replacement of a character, or characters, in the original simulated fragment. Both extra character rate and character deletion rate refer to characters being added to or removed from the simulated fragment. Vertical gap rate exemplifies the lack of entire columns in the fragment, which is a common occurrence in ancient papyrus collections. (See Table 4).

14,100 fragments were simulated where each fragment was changed based on a single key variable while the other variables were not used to alter the simulated fragment. Based on the varying key variables, fragments were categorized into the following five categories: unedited sequences, deletion rate sequences, extra character rate sequences, vertical gap rate sequences, and error rate sequences. For each of the 10 randomly selected manuscripts, five locations were randomly chosen as starting locations for simulated fragments. 15 different fragment lengths, ranging from 10 characters to 150 characters with a difference of 10 characters between subsequent lengths, were chosen for each fragment sequence. We used four possible values for error rate, deletion rate, and extra character rate: 0.0, 0.05, 0.1, and 0.2. Additionally, we used six possible values for the vertical gap rate: 0, 2, 4, 6, 8, and 10. Permutations of all possible combinations were aggregated and resulted in 13,500 (10 x 5 x

Table 5: Percentages of identification for each subset of fragment sequences used in the evaluation.

Subset	Number of Sequences	Identified	Ratio
Unedited	750	672	89.6%
Deletion Rate	2,950	2,612	88.5%
Ex. Char Rate	2,950	2,626	89.0%
Error Rate	3,000	2,678	89.2%
Vert. Gap Rate	4,450	3,871	86.9%

15 (4 + 4 + 4 + 6)) simulated fragment sequences. We removed 50 identical sequences from deletion rate, extra character rate, and vertical gap rate subsets. In addition to the modified sequences, 750 unmodified simulated sequences were aggregated into a pool of unedited sequences.

Results

As expected, the subset of unedited simulated fragment sequences that suffered no level of deterioration had the highest percentage of identification at 89.6% (See Table 5). The subsets containing fragment sequences modified based on error rate and extra character rate received the next best percentages of identification at 89.2% and 89.0% respectively, followed by the subset of sequences modified with deletion at an identification ratio of 88.5%. The worst performing subset was the subset of simulated fragment sequences that were modified based on the vertical gap rate.

For all subsets, the relative key variable became less significant as the length of fragment sequences became larger. While naturally apparent in all subsets, the pattern was particularly noticeable for the subset of fragment sequences that were modified based on vertical gap rate (see Figure 14). Sequences in this subset with a fragment length of 10 characters and any length of vertical gap were not identifiable. Once the length of the fragment was extended to 20 characters, the simulated fragment was identifiable despite being affected by a vertical

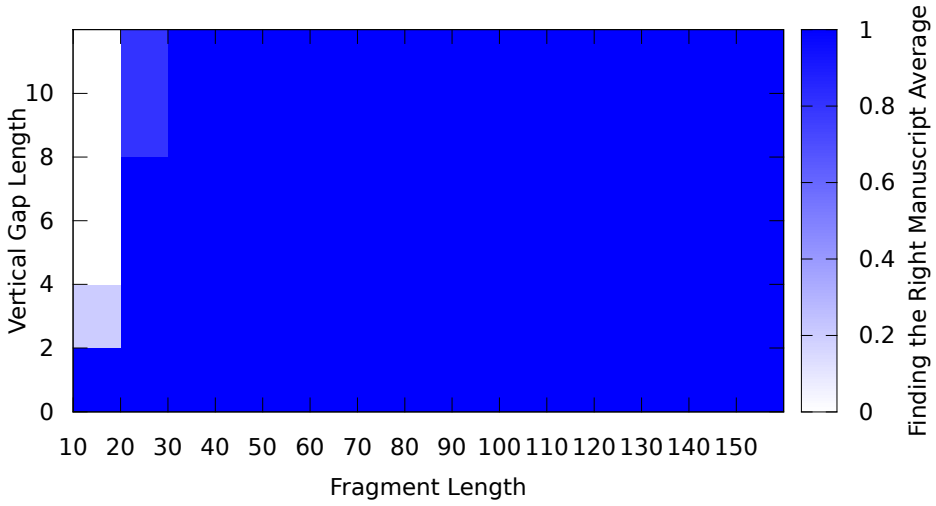

Figure 14: Above is a plot representing the average identification performance of all simulated fragments in the vertical gap rate subset based on a vertical gap rate ranging from 0 to 10 and fragment lengths ranging from 10 characters to 150 characters. Dark blue symbolizes the highest level of identification performance and white symbolizes the lowest level of identification performance.

gap.

Conclusion

In this paper, we observed a deficiency in the current rate of Ancient Greek papyri identification. Papyrologists try to manually match an 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 repetitious process of manual identification, we introduced a new methodology that aims to leverage genetic sequence alignment algorithms for Ancient Greek papyrus identification. With this methodology, we developed on Greek-BLAST, a BLAST variant for identifying and matching Ancient Greek text fragments. In a preliminary evaluation using an identity substitution matrix with a score of 10 on the diagonal and a score of -10 elsewhere, only 18

out of 8,956 simulated fragments were identified as the highest scoring match. Based on the presented evaluation, the calculation and integration of a new substitution matrix was a crucial step in the proposed methodology as Greek-BLAST outperforms other computational methods and tools used to quickly identify damaged, unknown fragments. Although we have chosen the BLAST algorithm to validate our approach, other genetic sequence alignment algorithms (*i.e.*, HMMER [16]) could take advantage of this methodology as well.

Future Work

Despite operating under the assumption that the consensus identification was the correct letter identification, Greek-BLAST was able to produce alignments with a high level of accuracy using simulated fragments. A key component of future work is to evaluate Greek-BLAST using more severely damaged fragments to identify key limitations of the GLOSUM matrix. Additional methodologies used to calculate empirical matrices from computational biology, such as the PAM (Accepted Point Mutation) matrix family [13], will be considered and investigated for application. Furthermore, in the future we plan on using an evaluation criterion that takes into account the entire retrieval, such as the Threshold Average Precision [11]. Once limitations of the matrix have been identified and resolved, we will perform a final assessment of the performance of Greek-BLAST using fragment transcriptions gathered from the Ancient Lives project that have already been matched to a known manuscript. Greek-BLAST's ability to identify papyri fragments from Oxyrhynchus would further validate both the usefulness of Greek-BLAST and the applicability of the proposed methodology.

CHAPTER V.

CONCLUSION

In papyrology, the tasks of transcribing and identifying papyrus fragments compose the foundation for understanding the culture of ancient civilizations. Due to the deteriorated nature of most papyrus fragments, these tasks are extremely slow and tedious.

There are currently two computational approaches for accelerating the task of transcription for deteriorated or torn documents, but each approach is unsuitable for transcribing most ancient texts, especially those that suffer from deterioration. The first approach is optical character recognition. A variety of optical character recognition methods have been developed for and applied to ancient texts and manuscripts. However, in many cases, these techniques fail due to quality and color inconsistencies in image datasets. Additionally, most OCR approaches have difficulty distinguishing between letters and tears in the document. The second approach and more favorable approach in recent research is crowdsourcing. Crowdsourcing overcomes the complications that come with an OCR-based approach by enlisting humans with the task of recognizing characters. However, most modern crowdsourcing transcription projects collect transcriptions using plain-text input fields. In many cases, users are incapable of transcribing all components of the document, especially content within the margin which might be the only information present. To overcome this obstacle, more recent applications, such as Ancient Lives, have replaced how users identify letters for transcriptions. In place of giving transcriptional information in plain-text input fields, users click on a letter in the image to create a marker with a x,y click-location and assign the marker a letter using an on-screen keyboard. While this approach does allow users to transcribe all components of a text, no method or technique presently exists that can make sense of transcriptional information gathered using this approach.

Unlike transcription, there has been no explicit effort that has aimed to accelerate the task of papyrus identification. However, it can be argued that this problem bears

similarity to fundamental problems in string similarity search and can thus share solutions. A number of approaches, such as the optimal matching algorithm, exist for identifying shared subsequences between strings or sequences. The most applicable of these approaches to the problem of identifying deteriorated papyrus fragments is genetic sequence alignment. By nature, amino acids that make up genetic sequences have the ability to mutate. In terms of computer science literature, a mutation is defined as an insertion, a deletion, or substitution of a character. A genetic sequence alignment between two genetic sequences finds the most longest common, biologically relevant subsequence shared between an anonymous sequence and any number of known genetic sequences. This process is identical to the task of identifying literary papyrus fragments as missing information in the papyrus (*e.g.*, from deterioration) are analogous to the concept of genetic mutations in terms of comparing sequence similarity. While the approach is very clearly applicable, no prior work has investigated how genetic sequence alignment can be usefully leveraged to the task of identification of deteriorated papyrus fragments.

Accelerated Transcriptions for Ancient Greek Papyrus Fragments

To allow transcriptional crowdsourcing applications to accommodate the data collection needs of ancient texts, a new computational pipeline has been developed to calculate accurate papyrus transcriptions from a large number (*e.g.*, millions) of crowdsourced letter identifications where each identification has an associated two-dimensional plane coordinate. The pipeline consists of two components. The first component contains two approaches, the stepwise approach and the kernel-based approach, for aggregating and clustering identifications into consensus identifications for each fragment. An evaluation of each implementation using professionally verified classifications suggested that the stepwise approach was slightly more effective for calculating consensus letter identifications for non-cursive, semi-cursive, and cursive fragments (see Figure 1). In addition, the stepwise approach is more practical than the kernel-based approach as the stepwise approach executes approximately 576 times

faster than the kernel-based approach.

Using the consensus identifications from the first component, the second component creates sequences (or strings) that represent digital transcriptions of the respective fragment. An evaluation of this component suggests that the component's underlying approach for calculating lines was effective for papyrus fragments with parallel lines. For 30 fragments with parallel lines, the edit distance between the calculated transcription and the known transcription was on average 15.6% of the relative fragment's length (see Figure 2). Conversely, the approach performed much better on papyrus fragments with curvilinear lines, in which the edit distance between calculated transcription and known transcription was on average 35.6% of the fragment's string length. The difference in performance might be explained by the choice in dataset. Only recently have papyrus fragment transcriptions been digitized, and to date, only 41 randomly selected fragments in the Oxyrhynchus papyri had been professionally prepared with digital transcriptions. A follow-up evaluation with additional data would offer more insight as to whether or not this evaluation is representative of the approach's true accuracy on papyrus fragments with curvilinear lines.

Accelerated Identification for Ancient Greek Papyrus Fragments

To accommodate the task of identification, a novel methodology is presented for leveraging genetic sequence alignment algorithms to the task of identifying an anonymous papyrus fragment to a known literary manuscript. A demonstration of this methodology has culminated in Greek-BLAST, a variant of the BLAST algorithm that can be used to identify deteriorated papyrus fragments to known literary manuscripts. In a formal evaluation using simulated fragments, Greek-BLAST identified 88.4% of the fragment queries as the highest-scoring identification. Nearly all cases of misidentification can be attributed to fragment length as simulated fragments with a length less than 10 characters were consistently unidentifiable. Once the length of the fragment had been extended to 20 characters, the fragment sequence was identifiable in spite of being modified severely by multiple key variables.

Greek-BLAST accelerates the process of papyrus identification from multiple days, weeks, or months to a few seconds for a single fragment. This is the only known work that aims to explicitly accelerate the task of identification for deteriorated papyrus fragments.

Future Work

The first task of future work is improving how transcription lines are created for papyrus fragments with curvilinear lines. Although effective for papyrus fragments containing strictly parallel lines of text, the proposed pipeline for calculating transcriptions yielded unsatisfactory results for papyrus fragments with curvilinear lines of text. The current implementation of the line sequencing component assumes that written lines of text are separated by a consistent height. For papyrus fragments with curvilinear lines of text, the height between lines of text can be variable, and in some cases, lines can overlap one another. Additionally, a follow-up evaluation with a larger test dataset will allow weaknesses to be more obviously exposed.

The second task of future work is examining Greek-BLAST's performance using transcriptions of real fragments that have already been identified to a known literary manuscript. Although success was evident with a large collection of simulated sequences, an evaluation with real data will reinforce the practicality of both the methodology and Greek-BLAST. Additional methodologies for calculating substitution matrices should be investigated for applicability and possibility of improving the accuracy of Greek-BLAST.

BIBLIOGRAPHY

[1] Abouelhoda, M. and Ghanem, M. String Mining in Bioinformatics. In *Scientific Data Mining and Knowledge Discovery*, pages 207–247. Springer, 2010.

[2] Altschul, S. F. Amino acid substitution matrices from an information theoretic perspective. *Journal of molecular biology*, 219(3):555–565, 1991.

[3] Altschul, S. F. and others,. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[4] Apostolico, A. and Giancarlo, R. Sequence alignment in molecular biology. *Journal of Computational Biology*, 5(2):173–196, 1998.

[5] Atkinson, Q. D. and Gray, R. D. Curious parallels and curious connections: phylogenetic thinking in biology and historical linguistics. *Systematic biology*, 54(4):513–526, 2005.

[6] Barzilay, R. and Lee, L. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 16–23. Association for Computational Linguistics, 2003.

[7] Berg-Kirkpatrick, T., Durrett, G., and Klein, D. Unsupervised transcription of historical documents. 2013.

[8] 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.

[9] Buechler, M., Heyer, G., and Gründer, S. eAQUA–bringing modern text mining approaches to two thousand years old ancient texts. In *Proceedings of e-Humanities–*

An Emerging Discipline, workshop at the 4th IEEE International Conference on e-Science, 2008.

[10] Buechler, M., Kruse, S., and Eckart, T. Bringing Modern Spell Checking Approaches to Ancient Texts - Automated Suggestions for Incomplete Words. In *Proceedings of Digital Humanities*, 2012.

[11] Carroll, H. D., Kann, M. G., Sheetlin, S. L., and Spouge, J. L. Threshold average precision (tap-k): a measure of retrieval designed for bioinformatics. *Bioinformatics*, 26(14):1708–1713, 2010.

[12] Chanda, S., Franke, K., and Pal, U. Document-zone classification in torn documents. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 25–30. IEEE, 2010.

[13] Dayhoff, M. and others.,. A model of evolutionary change in proteins. In *In Atlas of protein sequence and structure*. Citeseer, 1978.

[14] Diem, M., Kleber, F., and Sablatnig, R. Text classification and document layout analysis of paper fragments. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 854–858. IEEE, 2011.

[15] Diem, M. and Sablatnig, R. Recognizing characters of ancient manuscripts. In *IS&T/SPIE Electronic Imaging*, pages 753106–753106. International Society for Optics and Photonics, 2010.

[16] Finn, R. D., Clements, J., and Eddy, S. R. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl 2):W29–W37, 2011.

[17] Fukunaga, K. and Hostetler, L. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32–40, 1975.

[18] Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[19] Henikoff, S. and Henikoff, J. G. Performance evaluation of amino acid substitution matrices. *Proteins: Structure, Function, and Bioinformatics*, 17(1):49–61, 1993.

[20] Kittur, A., Chi, E., and Suh, B. Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements. *Proc. CHI 2008*, 2008.

[21] Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

[22] Kondrak, G. Phonetic alignment and similarity. *Computers and the Humanities*, 37(3):273–291, 2003.

[23] Levenshtein, V. I. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966.

[24] Lipman, D. J. and Pearson, W. R. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, 1985.

[25] MacLeod, R. *The Library of Alexandria: centre of learning in the Ancient World*. IB Tauris, 2005.

[26] Manning, C. D. and Schütze, H. *Foundations of statistical natural language processing*. MIT press, 1999.

[27] Mori, S., Nishida, H., and Yamada, H. *Optical character recognition*. John Wiley & Sons, Inc., 1999.

[28] Mount, D. W. Sequence and genome analysis. *Bioinformatics: Cold Spring Harbour Laboratory Press: Cold Spring Harbour*, 2, 2004.

[29] Moyle, M., Tonra, J., and Wallace, V. Manuscript transcription by crowdsourcing: Transcribe Bentham. *Liber Quarterly*, 20(3/4):347–356, 2011.

[30] Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[31] Pearson, W. R. and Lipman, D. J. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

[32] Prather, E. E., Cormier, S., Wallace, C. S., Lintott, C., Raddick, M. J., and Smith, A. Measuring the Conceptual Understandings of Citizen Scientists Participating in Zooniverse Projects: A First Approach. *Astronomy Education Review*, 12(1):010109, 2013.

[33] Prokić, J., Wieling, M., and Nerbonne, J. Multiple sequence alignments in linguistics. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 18–25. Association for Computational Linguistics, 2009.

[34] Smith, D. A., Rydberg-Cox, J. A., and Crane, G. R. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25, 2000.

[35] Smith, T. F. and Waterman, M. S. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[36] Vamvakas, G., Gatos, B., Stamatopoulos, N., and Perantonis, S. J. A complete optical character recognition methodology for historical documents. In *Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on*, pages 525–532. IEEE, 2008.

[37] Van Rijsbergen, C. J. Foundation of evaluation. *Journal of Documentation*, 30(4):365–373, 1974.

[38] Williams, A. C., Carroll, H. D., Wallin, J. F., Brusuelas, J., Fortson, L., Lamblin, A.-F., and Yu, H. Identification of ancient greek papyrus fragments using genetic sequence alignment algorithms. In *e-Science (e-Science), 2014 IEEE 10th International Conference on*, volume 2, pages 5–10. IEEE, 2014.

[39] Williams, A. C., Wallin, J. F., Yu, H., Perale, M., Carroll, H. D., Lamblin, A.-F., Fortson, L., Obbink, D., Lintott, C. J., and Brusuelas, J. H. A computational pipeline for crowdsourced transcriptions of ancient greek papyrus fragments. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 100–105. IEEE, 2014.