Identification of Ancient Greek Papyrus Fragments Using Genetic Sequence Alignment Algorithms

Alex C. Williams*†, Hyrum D. Carroll*, John F. Wallin*, James Brusuelas†, Lucy Fortson§, Anne-Francoise Lamblin§, and Haoyu Yu§
*Middle Tennessee State University, Murfreesboro, Tennessee, USA
{Alex.Williams, Hyrum.Carroll, John.Wallin}@mtsu.edu
†University of Oxford, Oxford, Oxfordshire, UK
James.Bruseulas@classics.ox.ac.uk
§University of Minnesota, Minneapolis, Minnesota, USA
fortson@physics.umn.edu, {lambl001,yuxxx084}@umn.edu

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

The identification of an anonymous string of characters with a collection of known text sequences is ubiquitous in computational biology. Genes are often represented by a sequence of continuous characters, each of which denotes an amino acid. Relationships are inferred by finding multi-letter patterns shared between the anonymous sequence and a known sequence. This process is commonly referred to as genetic sequence alignment.

In this paper, we introduce a novel methodology that uses modern genetic sequence alignment algorithms as a method for identifying Ancient Greek text fragments. This application will offer papyrologists and other professionals in the humanities the ability to rapidly identify severely damaged texts. This approach leverages a new form of non-contextual, multi-line text identification for the Greek language that can greatly accelerate the tedious task of transcription and identification.

Keywords*—*identification; genetic sequence alignment; Ancient Greek; papyrus

I. 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 1). 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 [1]. 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.

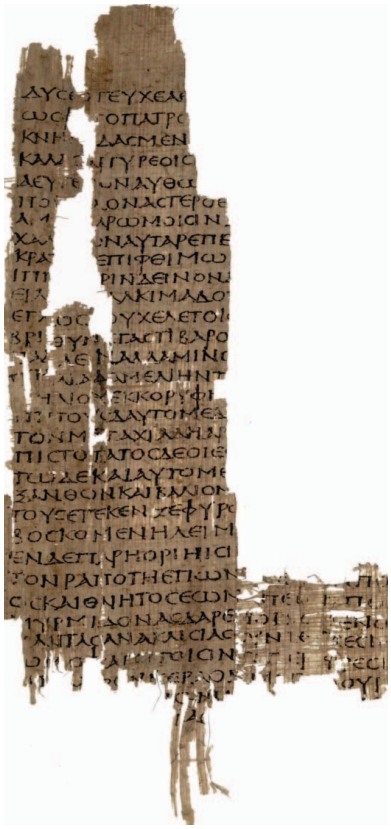

Fig. 1. A papyrus fragment from Oxyrhynchus. There are obvious gaps in the image caused by degradation of the papyri. In some cases, letters have been partially lost due to gaps. Notably, most of the text on the right side is missing. The hand of the scribe is extremely clear in this literary fragment, but in many cases, the handwriting is more difficult to read.

Past research has demonstrated that computational biology

978-1-4799-4287-9/14 $31.00 © 2014 IEEE
DOI 10.1109/eScience.2014.14

```
>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
```

Fig. 2. Example alignment of two genetic sequences by BLAST [2].

solutions can be usefully applied to data mining and machine learning problems [3]. 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 (see Figure 2). 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 as genetic sequence alignment. 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.

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.

II. Related Work

A. 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 [4] and historical linguistics [5]. Past work has demonstrated that homology search problems in computational biology can usefully take advantage of identical sequence alignment algorithms and techniques [6]. 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.

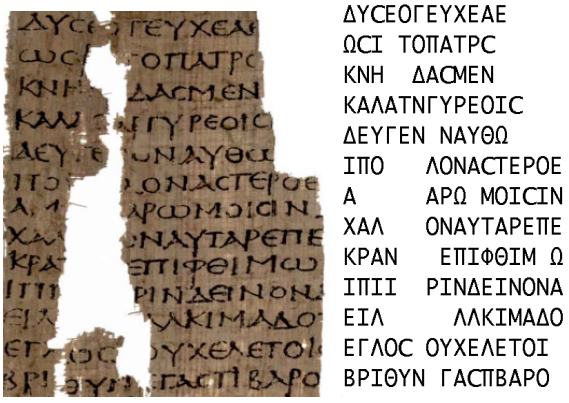

Fig. 3. A section of the original papyrus and the crowdsourced transcription. The crowdsourced transcription is shown immediately next to a portion of the original image of the papyrus from Figure 1. The consensus transcript is used for document identification.

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

Ancient Lives fragment: 131383

FRAGMENT ΠΑ?ΑΔΟ?Ι?ΑΓΑΘΗΚΑΙΠΑΝΔΩΡΗΜΑΤΕΛΕΙΟΝΑΝΩΘΕΝΕ?ΤΙΝΚΑΤΑΒΑ
TEXT     ΠΑΣΑΔΟΣΙΣΑΓΑΘΗΚΑΙΠΑΝΔΩΡΗΜΑΤΕΛΕΙΟΝΑΝΩΘΕΝΕΣΤΙ-ΚΑΤΑΒΑ
SIMILAR  ΠΑ ΑΔΟ Ι ΑΓΑΘΗΚΑΙΠΑΝΔΩΡΗΜΑΤΕΛΕΙΟΝΑΝΩΘΕΝΕ ΤΙ ΚΑΤΑΒΑ

FRAGMENT ΙΝΟΝΑΠΟΤΟΥΠΑΤΡΟ?ΤΩΝΦΩΤΩΝ
TEXT     ΙΝΟΝΑΠΟΤΟΥΠΑΤΡΟΣΤΩΝΦΩΤΩΝ
SIMILAR  ΙΝΟΝΑΠΟΤΟΥΠΑΤΡΟ ΤΩΝΦΩΤΩΝ
```

Fig. 4. 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.

B. 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 [7]. This project, like others in the Zooniverse, enlists volunteers to help process and analyze data [8]. 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 3). 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 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 damaged fragments, such as those found at Oxyrhynchus, content is presented with multiple incomplete or missing characters and words throughout the fragment.

III. 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) [2]. This new BLAST variant has been appropriately named Greek-BLAST.

A. 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 [11]. 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:

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

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 [12].

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

(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)	G (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)

TABLE I. THE GREEK LETTER ORIENTED SUBSTITUTION MATRIX (GLOSUM).

contained a score of -4 (see Table I).

IV. 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.

A. 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 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 II).

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 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.

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

TABLE II. EXAMPLE OF A SIMULATED FRAGMENT WITH KEY VARIABLE MODIFICATIONS HIGHLIGHTED IN RED.

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%

TABLE III. PERCENTAGES OF IDENTIFICATION FOR EACH SUBSET OF FRAGMENT SEQUENCES USED IN THE EVALUATION.

B. 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 III). 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 5). 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 gap.

V. 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 [13]) could take advantage of this methodology as well.

A. 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 [14], 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 [15]. 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.

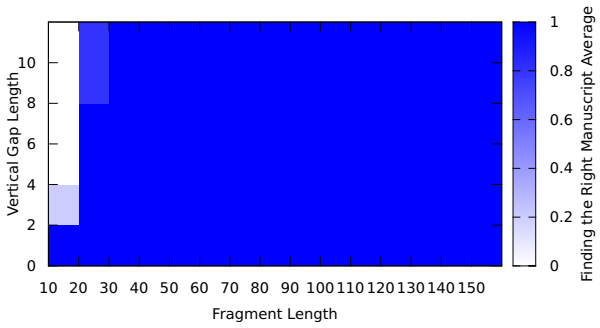

Fig. 5. A plot representing the average identification performance of all simulated fragments in the vertical gap rate subset based on a vertical gap rate ranging from 0 to 10 and fragment lengths ranging from 10 characters to 150 characters. Dark blue symbolizes the highest level of identification performance and white symbolizes the lowest level of identification performance.

ACKNOWLEDGMENTS

We would like to acknowledge Nita Krevans, Dirk Obbink, Marco Perale, Phillip Sellew, and Trevor Wennblom for their contributions to this project. Additionally, we would like to acknowledge the Egypt Exploration Society and the Ancient Lives project for access to the data set used in this study.

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

REFERENCES

[1] A. K. Bowman, R. A. Coles, N. Gonis, D. Obbink, and P. J. Parsons, *Oxyrhynchus: a City and its Texts*. Egypt Exploration Society, 2007, vol. 93.

[2] S. F. Altschul *et al.*, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[3] M. Abouelhoda and M. Ghanem, "String Mining in Bioinformatics," in *Scientific Data Mining and Knowledge Discovery*. Springer, 2010, pp. 207–247.

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

[5] J. Prokić, M. Wieling, and J. Nerbonne, "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*. Association for Computational Linguistics, 2009, pp. 18–25.

[6] Q. D. Atkinson and R. D. Gray, "Curious parallels and curious connections: phylogenetic thinking in biology and historical linguistics," *Systematic biology*, vol. 54, no. 4, pp. 513–526, 2005.

[7] "Ancient Lives," https://ancientlives.org.

[8] "Zooniverse," https://www.zooniverse.org.

[9] M. Buechler, G. Heyer, and S. Gründer, "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] M. Buechler, S. Kruse, and T. Eckart, "Bringing Modern Spell Checking Approaches to Ancient Texts - Automated Suggestions for Incomplete Words," in *Proceedings of Digital Humanities*, 2012.

[11] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10 915–10 919, 1992.

[12] ——, "Performance evaluation of amino acid substitution matrices," *Proteins: Structure, Function, and Bioinformatics*, vol. 17, no. 1, pp. 49–61, 1993.

[13] R. D. Finn, J. Clements, and S. R. Eddy, "Hmmer web server: interactive sequence similarity searching," *Nucleic acids research*, vol. 39, no. suppl 2, pp. W29–W37, 2011.

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

[15] H. D. Carroll, M. G. Kann, S. L. Sheetlin, and J. L. Spouge, "Threshold average precision (tap-k): a measure of retrieval designed for bioinformatics," *Bioinformatics*, vol. 26, no. 14, pp. 1708–1713, 2010.