Using concepts in natural language processing to determine the composition of artifact residues

Tung Tho Nguyen[1,] * and Korey J. Brownstein[2,] *

[1]The University of Chicago, Department of Mathematics, Chicago, Illinois, 60637, USA
[2]The University of Chicago, Department of Molecular Genetics and Cell Biology, Chicago, Illinois, 60637, USA

*Correspondence:
Tung Tho Nguyen
tungnt@uchicago.edu

Korey J. Brownstein
kbrownstein@uchicago.edu

Abstract

Determining the composition of artifact residues is a central problem in ancient residue metabolomics. Traditionally, this is done by comparing mass spectra features in common with an experimental pipe and a sample pipe (classical method). While this method is simple and straightforward, its prediction capabilities might be inaccurate. Here, we introduce a novel approach based on ideas from the field of natural language processing to solve this problem. We tested our strategy on a set of modern clay pipes. To limit biases, we were not provided information on which plant species had been smoked in which clay pipe. The results indicate that our algorithms performed 12.5% better than the previously published classical method.

Keywords: ancient residue metabolomics, archaeometry, artifact, mass spectral feature, smoking pipe

Introduction

Metabolomics is the systematic quantitative and qualitative study of small molecules (or mass spectral features) in biological systems. Brownstein *et al.* (2020) expanded upon this field with their ancient residue metabolomics-based method. Albeit the mass spectral features in this study were not derived from biological systems, they were residues left behind from biological processes; i.e., originating from plants including several *Nicotiana* species that were smoked by indigenous peoples. Before the Brownstein *et al.* (2020) study, ancient residue analysis relied on the biomarker approached. However, the biomarker approach failed to distinguish between related species, leaving open questions about the relationship between plants and people. For ancient residue metabolomics, all compounds are of interest improving the resolution of determining which plants species had been smoked in a particular pipe (Brownstein *et al.*, 2020).

In short, data from hyphenated chromatography instruments (such as gas chromatography- and liquid chromatography-mass spectrometer) are processed and aligned in MZmine 2 (Pluskal *et al.*, 2010), Progenesis QI (Waters Corporation, Milford, MA, USA), or another "omics" software. Afterwards, these data are exported from the software and then processed manually as described in the following passage from Brownstein *et al.* (2020): "[T]he dataset was exported into Microsoft Excel and mass spectral features shared with the blank [samples] were removed from the analysis. The [three solvent] extracts from each sample were combined into a single compound list and compounds with no abundance values were removed. The [ancient samples] were then compared to the [experimental samples using a Venn diagram]." This final component of processing and analyzing the datasets determines which plant species may have been used in an ancient artifact. Because this step requires a manual process, it can introduce errors and is time consuming. Various metabolomics platforms exist including MetaboAnalyst 4.0 (Pang *et al.*, 2021) or XCMS Online (Tautenhahn *et al.*, 2012); however, these platforms are limited in their ability to process datasets from ancient residue studies. Therefore, we introduce a novel, automated method for determining the composition of organic residues in modern smoking pipes. Our approach is inspired by techniques and ideas from the field of natural language processing (NLP).

Materials and Methods

A novel functionality of our approach is to introduce an entirely new method to compare the mass spectra feature similarities between experimental and ancient artifacts (i.e., sample groups). Our approach is inspired by advances in NLP (Cohen *et al.*, 2004; Cong *et al.*, 2017; Goodfellow *et al.*, 2016; Young *et al.*, 2018). Here we use the following analogy:

Words ←→ Mass Spectra Features
Documents ←→ Sample Groups

The standard technique in NLP is to first transform the original data into the term frequency-inverse document frequency (TF-IDF) matrix (Goodfellow *et al.*, 2016). This transformation helps to resolve the fact that some substances appear more often than others. More precisely, the importance of a term is not solely determined by its frequency in a text (TF) but also how rare this term is in other texts in the corpus (IDF). Let us recall these terminologies mathematically. Term frequency refers to the frequency of a word in a particular document:

$$tf(w,d) = \frac{\text{count of w in d}}{\text{number of words in d}}$$

The inverse of the document frequency which measures the informativeness/prevalence of term t

$$idf(w) = \log\left(\frac{N}{(df(w)+1)}\right),$$

where N is the number of documents and df(w) is the number of documents containing w. IDF score depends on the occurrence of terms and not on their numerical frequencies. Once the TF-IDF is computed, we can then use cosine similarity to compare two different groups. Recall that for two vectors v, w their cosine similarity is defined to be cosine of the angle θ between them, namely

$$\text{similarity} = \cos(\theta) = \frac{\langle v, w \rangle}{||v||||w||}$$

Here, $\langle v, w \rangle$ is the inner product of v, w and $||v||$,$||w||$ is the Euclidean norm of v, w. We note similarity score ranges from -1 meaning exactly opposite to 1 meaning exactly the same, with 0 indicating orthogonality, while in-between values indicate intermediate similarity or dissimilarity.

Artemisia ludoviciana Nutt. (*Alu*) leaves, *Arctostaphylos uva-ursi* (L.) Spreng. (*Auv*) leaves, *Cornus sericea* L. (*Cse*) bark, *Gaultheria shallon* Pursh (*Gsh*) leaves, *Lobelia inflata* L. (*Lin*) leaves, *Nicotiana attenuata* Torr. ex S. Watson (*Nat*) leaves, *Nicotiana glauca* Graham (*Ngl*) leaves, *Nicotiana obtusifolia* M. Martens & Galeotti (*Nob*) leaves, *Nicotiana quadrivalvis* Pursh (*Nqu*) leaves, *Nicotiana rustica* L. (*Nru*) leaves, *Nicotiana tabacum* L. (*Nta*) leaves, *Rhus glabra* L. (*Rgl*) autumn leaves, *Salvia sonomensis* Greene (*Sso*) leaves, *Taxus brevifolia* Nutt. (*Tbr*) needles, and *Verbascum thapsus* L. (*Vth*) leaves were collected, freeze-dried for 3 days, and crushed for experimental smoking. American Spirit (AmSp) tobacco (Santa Fe Natural Tobacco Company, Oxford, NC, USA) was purchased from a local grocery store in Pullman, Washington, USA.

The plant materials and AmSp were smoked following the experimental conditions detailed in Brownstein *et al.* (2020). To limit biases, the authors did not know which plant species had been smoked in which clay pipe. After the samples were analyzed by liquid chromatography-mass spectrometry and processed in MZmine 2 (Pluskal *et al.*, 2010) following the parameters described in Brownstein *et al.* (2020), the data were exported into .csv files. Python libraries, such as Sklearn and Pandas, were then used to apply the TF-IDF computation scores to these datasets.

Results and Discussion

We used Python to write the scripts because of the availability of several useful data analysis, machine learning, and deep learning libraries. All the scripts and datasets are freely available on GitHub: https://github.com/tungprime/NLP_and_composition_of_artifact_residues. Our script automates the classical method described in Brownstein *et al.* (2020), as well as utilizes recent advances in machine and deep learning to better predict which plant species had been smoked in a particular artifact (new method). As shown in Table 1, the new method predicts that CP1 was most likely smoked with *Nta* (0.0370). Table 2 summarizes the model predictions of the classical and new methods, and the key provides the expected results. In fact, CP1 was smoked with *Nta*. While the classical method only predicted four out of eight (50.0%) of the samples correctly, the new method performed slightly better, i.e., it classified five out of eight (62.5%) of the samples correctly (Table 2).

Table 1. Similarity scores of clay pipe 1 (CP1) smoked with an unknown plant sample. Sixteen (16) different experimental pipes smoked with only one of the plant species or AmSp were individually compared to CP1. The top five scores were only included in the table. *Nta* is the most likely candidate smoked in CP1. *Nat*, *Nicotiana attenuata*; *Ngl*, *Nicotiana glauca*; *Nob*, *Nicotiana obtusifolia*; *Nta*, *Nicotiana tabacum*; and AmSp, American Spirit.

	Nta	*Ngl*	*Nat*	AmSp	*Nob*
Similarity scores	0.0370	0.0287	0.0232	0.0185	0.0170

Contamination is a significant concern for ancient residue metabolomics (Damitio *et al.*, 2021; Zimmermann *et al.*, 2021). For instance, residues from commercial tobacco smoke may contaminate the surface of artifacts at excavation sites or on display at a museum. Thus, we included AmSp in our study as a contaminate control. With the contaminate control (0.0185), we were still able to accurately determine the composition of CP1 (Table 1) and the other clay pipes. Utilizing contaminate controls will improve confidence in whether or not a particular artifact had been smoked with an endemic tobacco. Furthermore, our new method will enable researchers to confidently determine if the caffeine present in/on an artifact resulted from ancient cacao or holly brewing practices instead of modern contaminates from caffeinated beverages such as coffee (King *et al.*, 2017; Washburn *et al.*, 2014).

Table 2. Predicted plant species in each clay pipe (CP). Sixteen (16) different experimental pipes smoked with only one of the plant species or AmSp were compared individually to each CP. The

expected results are the plant species listed under key. *Alu*, *Artemisia ludoviciana*; *Auv*, *Arctostaphylos uva-ursi*; *Cse*, *Cornus sericea*; *Lin*, *Lobelia inflata*; *Nat*, *Nicotiana attenuata*; *Ngl*, *Nicotiana glauca*; *Nob*, *Nicotiana obtusifolia*; *Nqu*, *Nicotiana quadrivalvis*; and *Nta*, *Nicotiana tabacum*.

Clay pipe (CP)	Classical method (number of mass spectra features shared with the experimental pipe)	Key	New method (similarity score)
CP1	*Auv* (3)	*Nta*	*Nta* (0.0370)
CP2	*Nat* (6) and *Nqu* (6)	*Nqu*	*Nqu* (0.1106)
CP3	*Nob* (10)	*Nob*	*Nob* (0.1145)
CP4	*Nat* (4)	*Alu*	*Nta* (0.0934)
CP5	*Ngl* (6)	*Ngl*	*Ngl* (0.0884)
CP6	*Lin* (10)	*Lin*	*Ngl* (0.0844)
CP7	*Auv* (13)	*Auv*	*Auv* (0.0735)
CP8	*Auv* (9) and *Cse* (4)	*Auv* and *Nta* mixture	*Cse* (0.0813) and *Auv* (0.0617)

It was also revealed that neither the new nor classical methods could accurately predict that CP8 had a mixture of *Auv* and *Nta* (Table 2). Both methods partially predicted the composition of CP8. Though the classical method performed slightly better because it had more hits for *Auv* than *Cse* (Table 2). Nonetheless, the experimental pipes compared to CP8 had only been smoked with *one* plant species. It is possible that training the new method with experimental pipes smoked with complex mixtures may improve the likelihood of predicting if a pipe had been smoked with more than one plant species.

Conclusion

Machine and deep learning have been vital tools for solving problems in biology where traditional methods seem inadequate or are time-consuming. Combining conventional and machine learning-based methods to process and analyze data helps researchers gain a more in-depth analysis of their data. The new method was used to predict which plant species had been smoked in modern clay pipes, and we believe this method can be applied to ancient smoking pipes, brewing vessels, and other artifacts.

Data availability statement: All scripts and datasets used in this study are in the GitHub repository: https://github.com/tungprime/NLP_and_composition_of_artifact_residues.

Author contributions: T.T.N. and K.J.B. designed the project. T.T.N. developed and wrote the scripts. T.T.N. and K.J.B. wrote the paper. All authors approved of the final version of the paper.

Funding: Korey Brownstein, Ph.D., holds a Postdoctoral Enrichment Award from the Burroughs Wellcome Fund. This research was also supported by National Science Foundation grant number 1906607.

Conflicts of interest: The authors declare no competing financial interests.

References

Brownstein, K. J., Tushingham, S., Damitio, W. J., Nguyen, T., and Gang, D. R. (2020). An ancient residue metabolomics-based method to distinguish use of closely related plant species in ancient pipes. *Frontiers in Molecular Biosciences* 7, 133.

Cohen, K. B., and Hunter, L. (2004). Natural language processing and systems biology (pp. 147–175). In: Dubitzky, W., and Pereira, F. (eds.), *Artificial intelligence and systems biology*. Springer, Dordrecht, Netherlands.

Cong, Y., Chan, Y. B., Phillips, C. A., Langston, M. A., and Ragan, M. A. (2017). Robust inference of genetic exchange communities from microbial genomes using TF-IDF. *Frontiers in Microbiology* 8, 21.

Damitio, W. J., Tushingham, S., Brownstein, K. J., Matson, R. G., and Gang, D. R. (2021). The evolution of smoking and intoxicant plant use in ancient Northwestern North America. *American Antiquity*, 1–19.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep Learning*. MIT Press, Cambridge, Massachusetts, USA.

King, A., Powis, T. G., Cheong, K. F., and Gaikwad, N. W. (2017). Cautionary tales on the identification of caffeinated beverages in North America. *Journal of Archaeological Science* 85, 30–40.

Pang, Z., Chong, J., Zhou, G., de Lima Morais, D. A., Chang, L., Barrette, M., Gauthier, C., Jacques, P. É., Li, S., and Xia, J. (2021). MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research* 49, W388–W396

Pluskal, T., Castillo, S., Villar-Briones, A., and Orešič, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11, 395.

Tautenhahn, R., Patti, G. J., Rinehart, D., and Siuzdak, G. (2012). XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical Chemistry* 84, 5035–5039.

Washburn, D. K., Washburn, W. N., Shipkova, P. A., and Pelleymounter, M. A. (2014). Chemical analysis of cacao residues in archaeological ceramics from North America: considerations of contamination, sample size and systematic controls. *Journal of Archaeological Science* 50, 191–207.

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 13, 55–75.

Zimmermann, M., Brownstein, K. J., Díaz, L. P., Aragón, I. A., Hutson, S., Kidder, B., Tushingham, S., and Gang, D. R. (2021). Metabolomics-based analysis of miniature flask contents identifies tobacco mixture use among the ancient Maya. *Scientific Reports* 11, 1–11.