analytical chemistry

Article

pubs.acs.org/ac

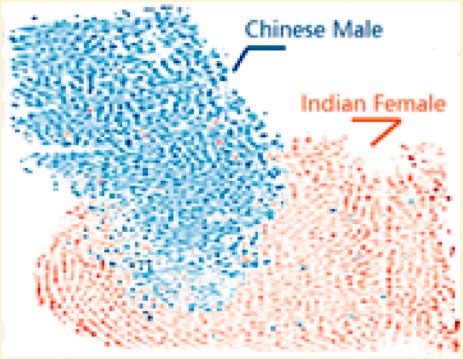
Personal Information from Latent Fingerprints Using Desorption Electrospray Ionization Mass Spectrometry and Machine Learning

Zhenpeng Zhou and Richard N. Zare*

Department of Chemistry, Stanford University, Stanford, California 94305-5080, United States

Supporting Information

ABSTRACT: Desorption electrospray ionization-mass spectrometry imaging (DESI-MSI) was applied to latent fingerprints to obtain not only spatial patterns but also chemical maps. Samples with similar lipid compositions as those of the fingerprints were collected by swiping a glass slide across the forehead of consenting adults. A machine learning model called gradient boosting tree ensemble (GDBT) was applied to the samples that allowed us to distinguish between different genders, ethnicities, and ages (within 10 years). The results from 194 samples showed accuracies of 89.2%, 82.4%, and 84.3%, respectively. Specific chemical species that were determined by the feature selection of GDBT were identified by tandem mass spectrometry. As a proof-of-concept, the machine learning model trained on the sample data was applied to overlaid latent fingerprints from different individuals, giving accurate gender and ethnicity information from those fingerprints. The results suggest that DESI-MSI imaging of fingerprints with GDBT analysis might offer a significant advance in forensic science.



Fingerprints are crucial in forensic sciences for identification of criminals.[1] Most fingerprint analysis methods focus on visual comparison and imaging. However, fingerprints, which are created mainly from sweat,[2,3] possess the potential to provide more personal information. Latent fingerprints are composed of the natural secretions of glands in the skin, principally eccrine and sebaceous glands. Eccrine sweat consists predominantly of water and a highly complex mixture of organic (e.g., amino acids, proteins, and lactate) and inorganic materials (e.g., Na^+, K^+, Cl^-, and trace metal ions).[4] Sebaceous secretions, called sebum, are predominantly composed of fatty acids, glycerides, cholesterol, squalene, and a variety of lipid esters. The chemical composition of sweat is known to differ between individuals but for any given individual to be essentially the same over the various parts of the body.[3,5] The molecules in sweat are products of metabolism, which are affected by several factors, including age, gender, and genetic inheritance.[6] Sweat, which is the main excretion in fingerprints,[7,8] is closely related to human metabolism.[9] Therefore, it is expected that its chemical analysis might offer personal information such as gender, age, ethnicity, medical history, and drug usage.

Mass spectrometry and spectroscopic imaging techniques has been applied to obtain fingerprint images.[10−16] Cooks and co-workers[17] as well as Perry and co-workers[11] have shown the capability of desorption electrospray ionization mass spectrometry imaging (DESI-MSI) for fingerprint imaging and explosives detection. These methods concentrate on obtaining visual patterns of fingerprints. Some work has also been devoted to chemical composition. Zhong and co-workers[18] developed laser-based mass spectrometry imaging for chemical maps of fingerprints that are found on a special film of semiconductors. Kazarian and co-workers[19,20] have performed studies in attenuated total reflection Fourier transform-infrared spectroscopy to obtain chemical maps of latent fingerprints. Francese and co-workers[21−24] have applied matrix-assisted laser desorption/ionization (MALDI) on chemical profiling of fingerprints and determination of genders. Dorrestein and co-workers[25] using chromatographic separations followed by mass spectrometric detection have correlated the chemicals on phones to people's lifestyles. We present an alternative mass spectrometric approach that we believe is much easier to implement.

The differences in lipid composition from various groups of people were studied and used for identification.[26−28] Halamek and co-workers[29,30] developed colorimetric methods for gender detection. Chilcott and co-workers[31] studied the effect of ethnicity, gender, and age on the amount and composition of sebum but found no differences. Skjold and co-workers[32] found out serum lipid concentrations in blood differ by gender and age.

In this work, ambient ionization mass spectrometry and machine learning were coupled to analyze latent fingerprints. The machine learning methods dug through the enormous amount of chemical information that mass spectrometry provided. In addition, by feature selection of the machine learning model and tandem mass spectrometry, the specific molecules that are different between individuals were pinpointed.

Received: November 15, 2016
Accepted: December 16, 2016
Published: January 5, 2017

DOI: 10.1021/acs.analchem.6b04498
Anal. Chem. 2017, 89, 1369−1372

MATERIALS AND METHODS

Human Subject Approval. The research was approved by Stanford Research Compliance Office's Human Subjects Research Institutional Review Board (IRB). The protocols were carried out in accordance with IRB regulations.

Fingerprint Collection. Eight study participants who are racially diverse and cover a span of ages washed their hands with soap and dried them in air, before placing their hands into polyethylene (PE) gloves for 60 min to accelerate perspiration. Samples for fingerprint imaging from each participant were produced by pressing his or her fingers onto a glass slide for 1 s.

Lipid Sample Collection. Lipid samples from fingerprints were collected by the procedure described above. Forehead lipid samples were collected from 203 study participants by swiping a glass slide across each of their foreheads.

Mass Spectrometry Imaging. Desorption electrospray Ionization (DESI) was set up for fingerprint imaging and lipid sample analysis. A custom-built DESI source with an $x-y$ stage coupled to an LTQ-Orbitrap XL mass spectrometer (Thermo Scientific) was used. The spectrum was collected under negative ion mode with m/z 150−1000. The DESI source used methanol−water (9:1 v/v) as the solvent with a flow rate of 1 μL/min. The nitrogen gas pressure was set to 80 psi. The spatial resolution of the imaging was estimated to be 200 μm.

Data Analysis. The Xcalibur raw files were read and converted into Python numpy (.npy) files. A hand-written peak finding algorithm converted the continuous spectrum to sparse peak data. A total of 1634 peaks were found in each sample. The data set was purged, which included discarding samples with too few peaks or low peak intensities, resulting in a sample size of 194. Each sample was then vectorized by the peak values with a resolution of 0.1 m/z. Samples were normalized by $l1$ norms of the sample vectors, which divided the sample vector by the sum of absolute values in the vector. Algorithms were adapted from xgboost[33] and scikit-learn.[34] The samples were separated into a training set, a cross-validation set, and a test set, with ratio of 7:1.5:1.5.

Classification algorithms of logistic regression, support vector machines, random forests, gradient tree boosting, nearest neighbors, and Bayesian regression were tested. Model selection was based on the performance of the cross-validation set.

RESULTS AND DISCUSSION

Different Source of Lipids. Lipids from foreheads and fingerprints were taken from eight people and analyzed by mass spectrometry. The spectra of lipids from the foreheads and fingers (Figure 1) showed no significant differences under statistical t test with 95% confidence interval. We concluded that different sources of lipids from the same people have similar compositions.[35]

Mass Spectrometry Imaging of Fingerprints. Figure 2 shows the representative negative-ion mode images from a fingerprint. Most of the species showed spatial homogeneity, indicating that the secretory products were nearly the same throughout the image. The spatial fingerprint pattern could be detected from the mass spectrometry imaging of the fingerprint, but our interest is in the chemical composition rather than the spatial distribution.

Tandem mass spectrometry was used to extract molecules from the fingerprints and determine the composition of each peak (Figure S1). Most of the ions detected in the mass spectra were identified as fatty acids (FA), tri(acyllalkyl)glycerols (TG), or di(acyllalkyl)glycerols (DG). Specifically, the peak at m/z = 227.20075 was identified to be FA(14:0), which represents all chain permutations of fatty acids with 14 carbons and 0 double bonds; the peak at m/z = 509.45715 was identified to be DG(16:0|12:1(OH)), which represents all chain permutations of diacylglycerols, whose two acyl chains are fatty acyls, one with 16 carbons and 0 double bonds and the other with 12 carbons with 1 double bond and 1 −OH substitution.

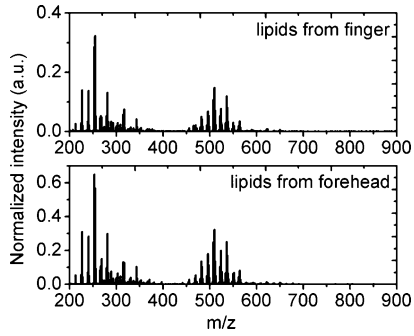


Figure 1. Different sources of lipids. The upper spectrum shows the lipids from a finger; the lower spectrum shows the lipids from a forehead of the same individual. They have no significant variance under the t test within 95% confidence interval.

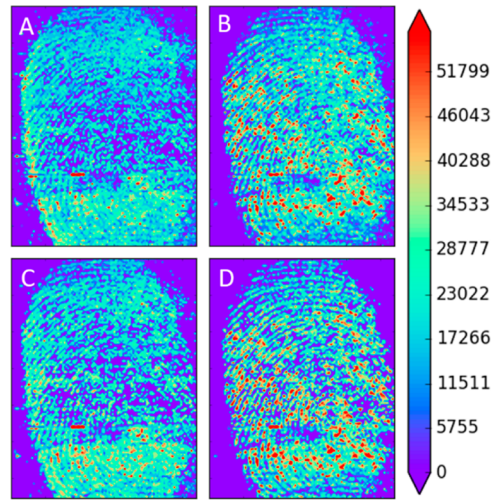


Figure 2. Selected negative-ion mode DESI imaging of the same fingerprint at (A) m/z = 227, (B) m/z = 241, (C) m/z = 253, and (D) m/z = 509. They show similar abundances across the fingerprint. Tandem mass spectrometry data shows that the four peaks can be (A) FA(14:0), (B) FA(15:0), (C) FA(16:1), and (D) DG(16:0|12:1(OH)). Abbreviations: FA is short for fatty acid, FA(14:0) represents all chain permutations of fatty acids with 14 carbons and 0 double bonds. DG is short for di(acyllalkyl)glycerols, DG(16:0|12:1(OH)) represents all chain permutations of diacylglycerols, whose two acyl chains are fatty acyls, one with 16 carbons and 0 double bonds and the other with 12 carbons with 1 double bond and 1 —OH substitution.

Classification by Machine Learning Models. By swiping a glass slide across the forehead of each study participants, samples with similar lipid compositions as that of the fingerprints were obtained. The total number of samples, from both fingerprints and foreheads, was 194. None of these represent replicates. A machine learning algorithm of gradient boosting tree ensemble (GDBT) (Supporting Information, Part I) was applied on the samples to classify them between different

DOI: 10.1021/acs.analchem.6b04498
Anal. Chem. 2017, 89, 1369−1372

genders, ethnicities (American, Chinese, European, and Indian), and ages (20, 30, 40−50, 60 and above). A discriminative model was trained on the training set, and the hyper-parameters were optimized on the cross-validation set. The final classification accuracy was 89.2%, 82.4%, and 84.3%, respectively, on test sets (Table 1), showing we could determine with good accuracy the gender, ethnicity, and age of a person from the lipid profile.

Table 1. Final Assessment on Test Sets[a]

	accuracy (%)	specificity (%)	precision (%)
gender	89.2	87.4	88.4
ethnicity	82.4	88.7	69.4
age	84.3	89.8	72.7

[a]The accuracy is defined as ((true positive + true negative)/total = (positive + negative)), specificity is defined as (true negative/(true negative + false positive)), and precision is defined as (true positive/(true positive + false positive)). The values shown are average values for multiclass classifications. These are believed to be the "standard" definitions in which specificity refers to how much of the set of data are true negative that are originally assigned as negative, and precision is how much of the set of data are true positive that are predicted by the model to be positive

Two overlaid fingerprints from different people were imaged by DESI-MSI to demonstrate the classification model. Lower resolution than Figure 2 was used to protect the privacy of individuals who provided their fingerprints. Figure 3A shows

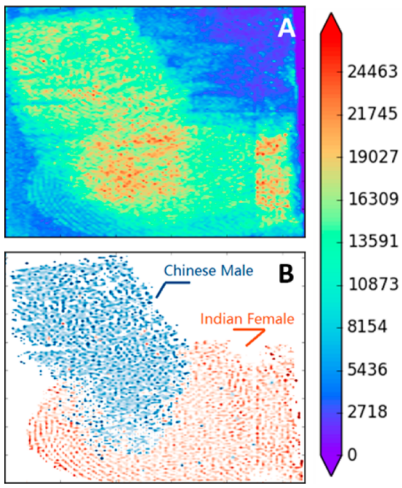


Figure 3. DESI-MSI and classification results of the fingerprint imaging: (A) negative-ion mode DESI-MS images of m/z = 253 and (B) the classification result of each pixel in the image by the pretrained model. The pixels predicted to be belong to a Chinese male are shown in blue, while the pixels predicted to be from an Indian female are shown in red. In both cases, the predictions were correct.

the negative ion mode DESI-MS ion images of m/z 253, in which the two fingerprints were recognizable. However, the boundaries of the fingerprints were not clear. We applied the pretrained model on each pixel of the mass spectrometry image and plot the classification result of each pixel in Figure 3B. The pixels predicted to be belong to a Chinese male are shown in blue, while the pixels to be from an Indian female are shown in red. The discriminative model was able get personal information from the fingerprints, resulting in a better separation of the fingerprints.

Feature Selection and Identification. The peak finding algorithm found 1634 peaks in the samples, indicating 1634 molecular features that could provide useful information. The large number of peaks makes data interpretation difficult. However, we only need to know the molecular differences in lipid profiles between different groups of people. The GDBT model was capable of feature selection by finding features that maximized the decreases of weighted impurity in a tree.[36] (Supporting Information, Part I) By ranking the features with their decreases of impurity in the model that yielded the lowest test errors, the relative feature importance in gender classification is shown in Figure 4A. Figure 4B shows the sample spectrum of male and female, zoomed at peak of m/z = 481.42, which was determined to be important by the feature-selection algorithm.

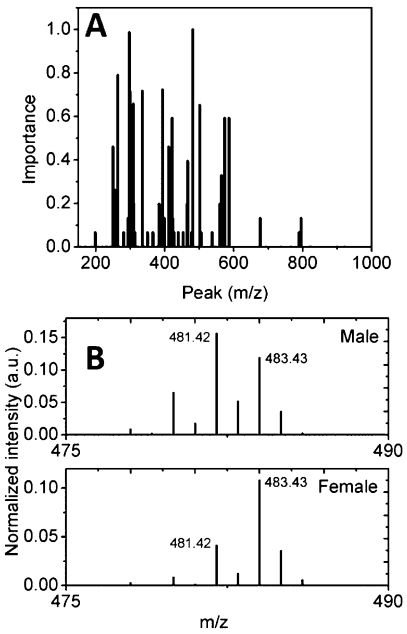


Figure 4. Feature selection and identification. (A) The relative importance of each peak in gender classification. (B) The sample spectrum of male and female, zoomed at peak m/z = 481.42 which is determined to be important for the feature selection algorithm.

Many peaks selected as important features by the model then were tentatively identified by tandem mass spectrometry with high mass resolution and accuracy. For example, the species with m/z = 481.42534 that was selected as an important feature in classification (with relative importance of 0.93) was then identified to be DG(16:1|10:0) (Figure S2A). Figures S3 and S4 show the features selected in the GDBT model to be important in the classification of ethnicity and age, as determined by tandem mass spectra of some important peaks in Figure S2B. The top 10 most important features (peaks) in classification of each category is shown in Table S1. The sample spectra of different ethnicities are shown in Figure S5. Although the chemical information on the features is not necessary for classification, the feature selection and identification results illustrate that the method could locate important molecules that reveal human metabolism variance between different groups.

CONCLUSION

In this work, mass spectrometry imaging was performed on fingerprints, from which pattern and chemical information were

DOI: 10.1021/acs.analchem.6b04498
Anal. Chem. 2017, 89, 1369−1372

captured at the same time. Personal information about gender, ethnicity, and age were obtained by applying a classification algorithm of gradient boosting tree ensemble on the lipid profiles from 194 samples, with accuracies of 89.2%, 82.4%, and 84.3%, respectively. The pretrained model was applied on two overlaid fingerprints, showing the capability of obtaining personal information about the two individuals from whom the fingerprints came from. In addition, by feature selection using the GDBT machine learning model, the species that were significant for classification between different groups of people were found and their chemical composition identified by tandem mass spectrometry. This information provides new chemical and biological insights into human metabolism. Finally, this work provides evidence that the mass spectrometry combined with machine learning can be a valuable tool for determining personal information by a noninvasive method.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b04498.

Brief introduction to gradient boosting tree; mass spectra; and top 10 most important features (peaks) in classification of gender, ethnicity, and age (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: zare@stanford.edu.

ORCID

Richard N. Zare: 0000-0001-5266-4253

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Support from the Air Force Office of Scientific Research under AFOSR Grant FA9550-16-1-0113 is gratefully acknowledged.

■ REFERENCES

(1) Payne, G.; Reedy, B.; Lennard, C.; Comber, B.; Exline, D.; Roux, C. *Forensic Sci. Int.* **2005**, *150* (1), 33−51.

(2) Asano, K. G.; Bayne, C. K.; Horsman, K. M.; Buchanan, M. V. *J. Forensic Sci.* **2002**, *47* (4), 805−807.

(3) Croxton, R. S.; Baron, M. G.; Butler, D.; Kent, T.; Sears, V. G. *Forensic Sci. Int.* **2010**, *199* (1−3), 93−102.

(4) Raiszadeh, M. M.; Ross, M. M.; Russo, P. S.; Schaepper, M. A.; Zhou, W. D.; Deng, J. H.; Ng, D.; Dickson, A.; Dickson, C.; Strom, M.; Osorio, C.; Soeprono, T.; Wulfkuhle, J. D.; Petricoin, E. F.; Liotta, L. A.; Kirsch, W. M. *J. Proteome Res.* **2012**, *11* (4), 2127−2139.

(5) Andres, P.; Zugaj, D.; Laffet, G.; Schoot, B. M.; Martel, P. *Personal Identification Based on Sebum Composition*. WO2011154259 A1, December 15, 2011.

(6) Hazarika, P.; Jickells, S. M.; Wolff, K.; Russell, D. A. *Angew. Chem., Int. Ed.* **2008**, *47* (52), 10167−10170.

(7) From the Analytical Scene Fingerprint Sweat Holds Sex Info. *Chem. Eng. News* **2015**, *93* (45), 29−32.10.1021/cen-09329-ad06

(8) Hier, S. W.; Cornbleet, T.; Bergeim, O. *J. Biol. Chem.* **1946**, *166* (1), 327−333.

(9) Jadoon, S.; Karim, S.; Akram, M. R.; Kalsoom Khan, A.; Zia, M. A.; Siddiqi, A. R.; Murtaza, G. *Int. J. Anal. Chem.* **2015**, *2015*, 164974.

(10) Bright, N. J.; Webb, R. P.; Bleay, S.; Hinder, S.; Ward, N. I.; Watts, J. F.; Kirkby, K. J.; Bailey, M. J. *Anal. Chem.* **2012**, *84* (9), 4083−4087.

(11) Comi, T. J.; Ryu, S. W.; Perry, R. H. *Anal. Chem.* **2016**, *88* (2), 1169−1175.

(12) Montalto, N. A.; Ojeda, J. J.; Jones, B. J. *Sci. Justice* **2013**, *53* (1), 2−7.

(13) Peng, T. H.; Qin, W. W.; Wang, K.; Shi, J.; Fan, C. H.; Li, D. *Anal. Chem.* **2015**, *87* (18), 9403−9407.

(14) Tang, H. W.; Lu, W.; Che, C. M.; Ng, K. M. *Anal. Chem.* **2010**, *82* (5), 1589−1593.

(15) Walton, B. L.; Verbeck, G. F. *Anal. Chem.* **2014**, *86* (16), 8114−8120.

(16) Yagnik, G. B.; Korte, A. R.; Lee, Y. J. *J. Mass Spectrom.* **2013**, *48* (1), 100−104.

(17) Ifa, D. R.; Manicke, N. E.; Dill, A. L.; Cooks, G. *Science* **2008**, *321* (5890), 805−805.

(18) Tang, X. M.; Huang, L. L.; Zhang, W. Y.; Zhong, H. Y. *Anal. Chem.* **2015**, *87* (5), 2693−2701.

(19) Ricci, C.; Phiriyavityopas, P.; Curum, N.; Chan, K. L. A.; Jickells, S.; Kazarian, S. G. *Appl. Spectrosc.* **2007**, *61* (5), 514−522.

(20) Ricci, C.; Kazarian, S. G. *Surf. Interface Anal.* **2010**, *42* (5), 386−392.

(21) Bailey, M. J.; Bradshaw, R.; Francese, S.; Salter, T. L.; Costa, C.; Ismail, M.; Webb, R. P.; Bosman, I.; Wolff, K.; de Puit, M. *Analyst* **2015**, *140* (18), 6254−6259.

(22) Bailey, M. J.; Bright, N. J.; Croxton, R. S.; Francese, S.; Ferguson, L. S.; Hinder, S.; Jickells, S.; Jones, B. J.; Jones, B. N.; Kazarian, S. G.; Ojeda, J. J.; Webb, R. P.; Wolstenholme, R.; Bleay, S. *Anal. Chem.* **2012**, *84* (20), 8514−8523.

(23) Francese, S.; Bradshaw, R.; Ferguson, L. S.; Wolstenholme, R.; Clench, M. R.; Bleay, S. *Analyst* **2013**, *138* (15), 4215−4228.

(24) Francese, S. *Cracking Crimes with Lasers*, TEDx, Sheffield Hallam University, 2015.

(25) Bouslimani, A.; Melnik, A. V.; Xu, Z.; Amir, A.; da Silva, R. R.; Wang, M.; Bandeira, N.; Alexandrov, T.; Knight, R.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (48), E7645−E7654.

(26) Kiens, B. *Chem. Phys. Lipids* **2008**, *154*, S11−S11.

(27) Man, M. Q.; Xin, S. J.; Song, S. P.; Cho, S. Y.; Zhang, X. J.; Tu, C. X.; Feingold, K. R.; Elias, P. M. *Skin Pharmacol Phys.* **2009**, *22* (4), 190−199.

(28) Singh, R.; Sharma, S.; Singh, R. K.; Mahdi, A. A.; Singh, R. K.; Gierke, C. L.; Cornelissen, G. *Clin. Chim. Acta* **2016**, *459*, 10−18.

(29) Brunelle, E.; Huynh, C.; Le, A. M.; Halamkova, L.; Agudelo, J.; Halamek, J. *Anal. Chem.* **2016**, *88* (4), 2413−2420.

(30) Huynh, C.; Brunelle, E.; Halamkova, L.; Agudelo, J.; Halamek, J. *Anal. Chem.* **2015**, *87* (22), 11531−11536.

(31) Shetage, S. S.; Traynor, M. J.; Brown, M. B.; Raji, M.; Graham-Kalio, D.; Chilcott, R. P. *Skin Res. Technol.* **2014**, *20* (1), 97−107.

(32) Grimsgaard, S.; Eggen, A. E.; Njolstad, I.; Lochen, M. L.; Skjold, F. *Circulation* **2005**, *111* (4), E53−E54.

(33) Chen, T.; Guestrin, C. *arXiv Preprint* **2016**, arXiv:1603.02754.

(34) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. *Journal of Machine Learning Research* **2011**, *12* (Oct), 2825−2830.

(35) Hadorn, B.; Hanimann, F.; Anders, P.; Curtius, H. C.; Halverson, R. *Nature* **1967**, *215* (5099), 416.

(36) Liaw, A.; Wiener, M. *R News* **2002**, *2* (3), 18−22.

DOI: 10.1021/acs.analchem.6b04498
Anal. Chem. 2017, 89, 1369−1372