Proc. Fla. State Hort. Soc. 119:383-387. 2006.

USING A MASS SPECTROMETER LIBRARY MATCHING SYSTEM TO IDENTIFY CITRUS AND OTHER FOOD/NON-FOOD PRODUCTS

KEVIN L. GOODNER[1]* AND VANESSA R. KINTON[2]
[1]*USDA/ARS Citrus & Subtropical Products Laboratory*
600 Avenue S, NW
Winter Haven, FL 33881

[2]*Alcohol & Tobacco Tax and Trade Bureau*
6000 Ammendale Road
Ammendale, MD 20705

Additional index words. chemsensor, citrus, MS, multivariate, SIMCA

Abstract. **A method that identifies products based on a composite mass spectrum using standard chemical library searching functions is presented. Composite mass spectra were collected by sampling the headspace of a product directly without separation prior to analysis by a mass spectrometer. A library of spectra for 51 products (5 soaps, 2 hand lotions, 4 potato chips, 4 ketchups, 2 peanut butters, 4 breath mints/gums, 13 citrus juices, 1 bourbon, 3 onions, 5 colas, 3 coffees, 5 peppers) was generated, and 7 unknowns samples (17 runs total with replicates) were tested against the library. Eleven of the 17 unknown sample runs were correctly identified with the top rated library match, four were identified as the second best match, and 2 were not identified in the top two matches. This level of correct matching (15 of 17 as best or second best match) is encouraging, suggesting that this technique could be used on a larger scale for product identification. This technique requires fewer analyses, doesn't require advanced statistical knowledge, and uses widely known mass spectral library tools. A SIMCA model identified all 9 citrus product samples in a validation data set.**

The concept and application of using measurements of an unknown item and identifying it by comparison to a library is not new. Some widely known uses are fingerprinting (Faulds, 1880), facial recognition (Midorikawa, 1998), and DNA analysis (Ballantyne et al., 1989). This technique of library searching has also been used to identify strains of microorganisms by ion mobility spectrometry (IMS) (Vinopal et al., 2002). By comparison of the complex plasmagrams, multiple microorganisms could be positively identified. It is not a robust technology since limited sample size makes analysis of false positives and false negatives impossible to determine. However, this technique shows great potential in decreasing the amount of time involved with accurate identification of the organisms.

One of the most common and well known applications of library searching in analytical chemistry is mass spectral libraries for use with mass spectrometers and gas chromatography-mass spectrometry (GC-MS) systems (Adams, 1995; McLafferty, 2000; Stein et al., 1998). These libraries work by utilizing the fragmentation patterns molecules exhibit when detected by mass spectrometry. These patterns are reproducible and are dependant upon the compound's structure, allowing a database to be formed which allows for unknowns to be identified by their representative fragmentation pattern. In fact, the incorporation of MS libraries helped drive adoption of GC-MS by both researchers and industry by enabling the separation (GC) and identification (MS) of individual compounds in a complex mixture. The two most common search methods are the INCOS (used in the NIST database/search) and PBM (used in the Wiley database/search) which use different algorithms to determine which library spectrum is the best match for that of the unknown (McLafferty et al., 1998; Stein, 1994; Stein and Scott, 1994).

This paper examines the potential of using a mass spectrometer for identifying commercial food items by their mass spectra without any chromatographic separation and employing mass spectral library search routines. The goal is to demonstrate that this technique can also be applied to a complex product using a MS-based chemical sensor utilizing headspace sampling. There are several applications for this work. One is in the field of forensics, when identification of an unknown substance is often necessary. Another is when a company is trying to match a competitor's product (e.g., orange juice). The competitor's product could be analyzed and searched against the in-house product library, and the top library matches would be the likely starting point to formulate a similar product. Additionally, this method requires only a single data point for each product as compared to multivariate techniques which require many replications per product which translates into a tremendous time savings over more cumbersome techniques.

Material and Methods

Fifty-one products (detailed in Table 1) were collected from a local grocery store and analyzed within 48 h. Approximately 1 g of solid samples (sliced) and 1 mL of liquid samples were analyzed.

The chemsensor used was a Gerstel ChemSensor 4440A that includes a headspace sampling unit (7694, Agilent Technologies, Palo Alto, Calif.) with a mass selective detector (MSD) (5973, Agilent Technologies, Palo Alto, Calif.). The instrument was used in the scan mode from 35 to 250 m/z for 1.2 min. The headspace sampling unit had a cycle time of 3.7 min, with a vial equilibrium time of 20 min at 80°C (176°F). The loop used a 3 mL-silcosteel-loop and the fill time was 0.15 min at 95°C (203°F). The loop equilibration time was 0.05 min and the sample was injected for 1.0 min. The transfer line temperature was constant at 105°C (221°F).

Compounds present in the headspace of all the samples were directly transferred and detected in the MSD. Spectra were averaged over the whole chemsensor run and then exported to NIST MS Search version 1.6 (Stein et al., 1998) which was used for building and searching the library. For SIMCA (Soft Independent Modeling of Class Analogy), the intensities of the ions present were summed and recorded us-

Mention of a trademark or proprietary product is for identification only and does not imply a guarantee or warranty of the product by the U.S. Department of Agriculture. The U.S. Department of Agriculture prohibits discrimination in all its programs and activities on the basis of race, color, national origin, gender, religion, age, disability, political beliefs, sexual orientation, and marital or family status.

*Corresponding author; e-mail: goodner@citrus.usda.gov

Table 1. Products used to generate library and their respective trademark owners.

Number	Product	Trademark owner
1	Dial soap	The Dial Corporation
2	Ivory soap	Proctor & Gamble
3	Irish Spring soap	Colgate-Palmolive Company
4	Suave soap	Unilever
5	Zest soap	Proctor & Gamble
6	Caress hand lotion	Cheeseborough-Pond's, Inc.
7	Body Shop hand lotion	The Body Shop International
8	Herr's sour cream & onion potato chips	Herr's Corp.
9	Utz sour cream & onion potato chips	UTZ Quality Foods, Inc.
10	Lay's sour cream & onion potato chips	Frito-Lay, Inc.
11	Pringles sour cream & onion potato chips	Proctor & Gamble
12	Hunts ketchup	ConAgra Foods, Inc
13	Heinz purple ketchup	H.J. Heinz Company.
14	Heinz red ketchup	H.J. Heinz Company.
15	Heinz regular ketchup	H.J. Heinz Company.
16	Jif peanut butter	The J.M. Smucker Co.
17	America's Choice peanut butter	The Great Atlantic & Pacific Tea Company, Inc.
18	Cert's peppermint breath mint	Warner-Lambert Company
19	Dentyne Ice chewing gum—peppermint	Warner-Lambert Company
20	Lifesaver's peppermint breath mint	Life Savers Corporation
21	Breathsaver's peppermint breath mint	Kraft Foods Holdings, Inc.
22	Tropicana NFC OJ with "lots of pulp"	Tropicana Products, Inc.
23	Tropicana NFC OJ with "some pulp"	Tropicana Products, Inc.
24	Tropicana NFC OJ with "no pulp"	Tropicana Products, Inc.
25	Tropicana NFC OJ with low acid	Tropicana Products, Inc.
26	Tropicana NFC OJ with double Vitamin C&E	Tropicana Products, Inc.
27	Tropicana NFC white grapefruit juice	Tropicana Products, Inc.
28	Sunny Delight	Proctor & Gamble
29	Florida's Natural NFC OJ with "lots of pulp"	Florida's Natural Growers
30	Florida's Natural NFC OJ with "some pulp"	Florida's Natural Growers
31	Florida's Natural NFC OJ with "no pulp"	Florida's Natural Growers
32	America's Choice reconstituted OJ	The Great Atlantic & Pacific Tea Company, Inc.
33	Minute Maid reconstituted OJ	The Coca-Cola Company
34	Simply Orange NFC OJ with pulp	The Coca-Cola Company
35	Kentucky Bourbon	None
36	White onion	None
37	Yellow onion	None
38	Purple onion	None
39	Coca-Cola	The Coca-Cola Company
40	Diet Coke	The Coca-Cola Company
41	Pepsi	Pepsico, Inc.
42	Diet Pepsi	Pepsico, Inc.
43	RC Cola	Cadbury Schweppes plc.
44	Folger's 100% Columbian Coffee	Proctor & Gamble
45	Folger's Classic Roast Coffee	Proctor & Gamble
46	Folger's French Roast Coffee	Proctor & Gamble
47	Habanero pepper	None
48	Jalapeño pepper	None
49	Green pepper	None
50	Poblano pepper	None
51	New Mexico Pepper	None

ing special macros (Gerstel, Baltimore, Md.). The resulting data for the SIMCA model consisted of a composite mass spectrum of all headspace components of the sample. SIMCA was performed using Pirouette 3.11 (Infometrix, Woodinville, Wash.) for the citrus samples.

Results and Discussion

A spectrum library was created using the NIST MS search software for each product. This library was then used for identifying unknown samples to determine the accuracy of the method. Two types of identification were investigated. The first determined if the library would correctly identify the class of product. This was tested by using various samples and looking at the matches produced by the searching software. The unknowns were chosen from four different types of products, additionally two of the unknowns were modified. The hand soap (#2) was used approximately 50% for washing hands and then tested. The sample chosen for peppermint had not been used in generating the library, but was a similar product to ones that had been used. The results are presented in Table 2. The table lists the product number, results of

Table 2. Various samples and the results from the library searching program.

Sample #	Factors for top hit			Factors for correct hit			Hit #	Qualitative match
	Fit	rFit	Prob	Fit	rFit	Prob		
2 (used)[z]	682	753	23.0				1	√
2 (used)[z]	714	765	72.9	677	726	17.7	2	√
Mint[y]	900	900	64.2				1	√
Mint[y]	894	894	62.6				1	√
7	865	867	97.9				1	√
7	863	864	98.0				1	√
12	849	903	21.3	836	838	12.8	5	X
12	756	798	17.2	755	767	16.5	4	≈

"Factors for correct hit" are only listed for those that were not the top hit.
[z]Used approximately 50% to wash hands over a 1-2 week period.
[y]Not in library, but similar: Sugarfree peppermint from Target (copyright Target Brands, Inc.).

the searching, and a qualitative judgment. The factors listed are the fit, reverse fit, and probability. The fit is a measure that ranges from 1 to 999 and indicates how well the spectrum being searched matches the library spectra, assuming it is a pure compound. The reverse fit (rFit) is a measure that also ranges from 1 to 999 and indicates how well the spectrum fits ignoring any non-matching peaks as impurities. The probability (Prob) indicates the likelihood that the library spectrum and the unknown spectrum are of the same compound assuming the unknown is contained in the library.

Table 2 lists all the factors for the highest rated match and the factors for the correct match. Additionally, it is noted what position the correct match was in the list of possible matches. In this test, a soap (sample #2), mint, lotion (sample #7), and ketchup (sample #12) were analyzed to see how well they were identified as a soap, mint, lotion, and ketchup. In order to test the flexibility of the method, a used bar of a soap (to simulate the effects of use and storage) that is in the library and a mint that was not in the library were analyzed. The qualitative judgments are the authors' estimations of how well this method worked. Three levels of success were used: √ for successful, ≈ for somewhat successful, and X for not very successful. The first sample (#2) was hand soap that had been used. The library produced the correct item for the best match (√) for the first replication, while with the second replication the top match was for a similar product (√) but correct match was second best hit (√). The peppermint tested was not in the library, but for both samples, the top returned match was of a peppermint from the library which is very promising for this application in that it identified the product group. The sample analyzed was a store-brand peppermint and both replications had very high scores for matching a name-brand peppermint with the second best scores significantly lower. This suggests that the store-brand is either a repackaged name-brand product, or designed to mimic the name-brand product. The hand lotion returned the correct product as the highest rated match for both replications and therefore was rated with a √, and the top matches were much better than the rest. The probability was approximately 98 for the hand lotion and less than 1 for the rest of the matches. The last product tested was ketchup. One replication was not correctly identified as either the top match or as the top match to ketchup and, therefore, was judged to be not-successful (X). The second ketchup replication had fewer samples between the top match and the correct match, which was 4th best match, and all the samples that had better scores were ketchup so it was given a ≈. Ketchup was a particularly difficult sample due to the low number of MS peaks and homogeneity among products. This demonstrates that this technique will not likely be very beneficial to samples that produce few MS peaks.

The second kind of identification examined was how well the library could identify a specific product as a subset of its class. For this aspect, citrus juice was chosen as the class. Citrus juice was chosen for several reasons. Many products are quite similar but sold from several producers. Second, the difficulty with citrus juices is the fact that it is a natural product and, therefore, is susceptible to seasonal variations, unlike a manufactured item. The results of the library tests are detailed in Table 3. The layout of the table is the same as Table 2. Three

Table 3. Various citrus samples and the results from the library searching program.

Sample #	Factors for top hit			Factors for correct hit			Hit #	Qualitative match
	Fit	rFit	Prob	Fit	rFit	Prob		
22	933	933	25.8				1	√
22	925	925	23.7	922	922	21.0	2	≈
22	933	933	32.5				1	√
24[z]	935	935	23.4				1	√
24[z]	933	933	22.5				1	√
24[z]	931	931	24.6	928	928	21.7	2	≈
27	919	919	38.8				1	√
27	901	901	27.1				1	√
27	908	910	25.2	908	908	25.2	2	≈

[z]Different lot number.

Table 4. SIMCA interclass distances for the 13 different citrus samples.

	Product numbers												
	22	23	24	25	26	27	28	29	30	31	32	33	34
22	0.0	1.1	6.2	6.7	14.7	9.6	19.1	3.8	6.9	8.0	20.0	15.8	13.3
23	1.1	0.0	2.9	2.3	8.7	5.0	15.8	2.8	2.8	6.1	11.4	10.8	7.1
24	6.2	2.9	0.0	1.6	5.8	12.5	17.4	6.0	4.9	8.2	15.1	12.7	6.6
25	6.7	2.3	1.6	0.0	7.6	10.8	23.5	5.2	4.5	10.0	19.6	13.4	8.7
26	14.7	8.7	5.8	7.6	0.0	15.6	24.8	9.8	9.0	14.1	10.9	10.1	6.8
27	9.6	5.0	12.5	10.8	15.6	0.0	14.9	5.6	11.5	11.5	25.6	18.2	14.8
28	19.1	15.8	17.4	23.5	24.8	14.9	0.0	13.4	18.3	13.5	26.7	21.5	18.4
29	3.8	2.8	6.0	5.2	9.8	5.6	13.4	0.0	2.2	4.9	10.3	7.7	5.0
30	6.9	2.8	4.9	4.5	9.0	11.5	18.3	2.2	0.0	9.0	9.5	6.0	4.7
31	8.0	6.1	8.2	10.0	14.1	11.5	13.5	4.9	9.0	0.0	22.7	16.0	11.6
32	20.0	11.4	15.1	19.6	10.9	25.6	26.7	10.3	9.5	22.7	0.0	3.5	8.1
33	15.8	10.8	12.7	13.4	10.1	18.2	21.5	7.7	6.0	16.0	3.5	0.0	3.5
34	13.3	7.1	6.6	8.7	6.8	14.8	18.4	5.0	4.7	11.6	8.1	3.5	0.0

Distances less than 3.0 are not considered well separated and are in bold typeface.

samples were chosen: an orange juice with "lots of pulp", an orange juice with no pulp, and a white grapefruit juice. The results are quite promising. Of the nine analyses, six were correctly identified with the highest rated match from the library (√). In each of the three analyses where the best library match was not correct, the second best match that was correct (≈). Additionally, in each of the three cases, the differences between the fit, reverse fit, and probability were very slight and in some cases identical.

A SIMCA model was constructed with the citrus samples' mass spectral fingerprints. This classification model correctly identified all 9 samples in the validation data set. This is better than the results obtained with the library matching method, but has the drawback of being more involved, requiring more data points to build the model, and requiring extensive statistical knowledge. A good diagnostic for a SIMCA model is the interclass distances which are basically the Euclidian distance between the center of each group where large interclass distances imply well separated classes. Table 4 shows the interclass distances for all the 13 different citrus samples. As a rule of thumb, classes with interclass distances larger than three are considered well separated (Kvalheim and Karstang, 1992). As seen in Table 4, few samples had interclass distances less than three.

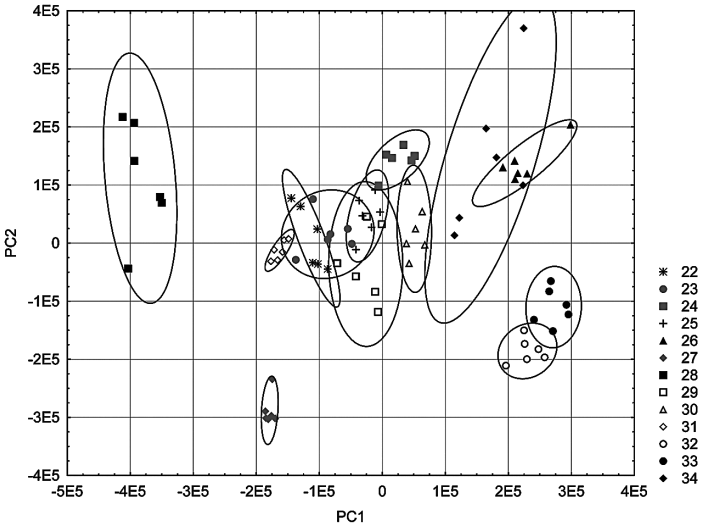

Fig. 1. Projections of the citrus samples into the space of the first two principal components. The ellipses do not provide statistical information and are provided for easier visual identification of the different clusters.

Another SIMCA model diagnostic is a class projection plot, shown in Fig. 1. This plot is similar to a PCA scores plot. In this study, the SIMCA model was constructed using a probability threshold of 0.95. It can be seen in Fig. 1 that for certain types of citrus samples the degree of separation is low (their ellipses overlap). The ellipses do not provide statistical information, but are provided for easier visual identification of the different clusters, and are not part of the SIMCA analysis. This implies that there are few differences in their mass spectral fingerprints. It can be possible that sampling the headspace of the citrus juice products does not provide sufficient analytes to completely discriminate between them. More validation data may be needed to test the robustness of this SIMCA model. Nevertheless, Sunny Delight (28), Tropicana NFC white grapefruit juice (27), and two reconstituted juices (32 and 33) were well separated from each other and the NFC orange juice products.

These results demonstrate that it is possible to utilize a MS chemical library searching program to search for and correctly identify products if their entire headspace is used to generate the spectra and the library.

Literature Cited

Adams, R. P. 1995. Identification of Essential Oil components by gas chromatography / mass spectroscopy; Allured Publishing Corp., Carol Stream, IL.

Ballantyne, J., G. Sensabaugh, and J. Witkowski. 1989. DNA technology and forensic science. Cold Spring Harbor Lab, Cold Spring Harbor, NY.

Faulds, H. 1880. On the skin-furrows of the hand. Nature 22:605.

Kvalheim, O. M. and T. V. Karstang. 1992. SIMCA-Classification by Means of Disjoint Cross Validated Principal Component Models, p. 237. In Multivariate pattern recognition in chemometrics, illustrated by case studies, R. G. Brereton (ed.). Elsevier, Amsterdam.

McLafferty, F. W., M. Y. Zhang, D. B. Staluffer, and S. Y. Loh. 1998. Comparison of Algorithms and Databases for Matching Unknown Mass Spectra. J. Amer. Soc. Mass Spectrom. 9:92-95.

McLafferty, F. W. 2000. Wiley Registry of Mass Spectral Data, Seventh Edition Database. Wiley, New York.

Midorikawa, H. 1998. The face pattern identification by back-propagation learning procedure. Neural Networks 1:515.

Stein, S. E. 1994. Estimating Probabilities of Correct Identification from Results of Mass Spectral Library Searches. J. Amer. Soc. Mass Spectrom. 5:316-323.

Stein, S. E. and D. R. Scott. 1994. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. J. Amer. Soc. Mass Spectrom. 5:859-866.

Stein, S. E. (Director), A. Mikaya, P. J. Ausloos, C. Clifton, S. G. Lias, V. Zaikin, and D. Zhu (Evaluators). 1998. NIST / EPA / NIH Mass Spectral Library—NIST 98 Version National Institute of Standards and Technology, Gaithersburg, MD.

Vinopal, R. T., J. R., Jadamec, P. deFur, A. L. Demars, S. Jakubielski, C. Green, C. P. Anderson, J. E. Dugas and R. F. DeBono. 2002. Fingerprinting bacterial strains using ion mobility spectrometry. Analytica Chimica Acta. 457:83-95.