COFFEE CHEMOMETRICS AS A NEW CONCEPT: UNTARGETED METABOLIC PROFILING OF COFFEE

C. LINDINGER[1], R.C.H. de Vos[2], C. Lambot[3], P. Pollien[1], A. Rytz[1], E. Voirol-Baliguet[1], R. Fumeaux[1], F. Robert[1], C. Yeretzian[4], and I. Blank[5]

[1] *Nestlé Research Center, Vers-chez-les-Blanc, 1000 Lausanne 26, Switzerland*
[2] *Plant Research International, 6700 AA Wageningen, The Netherlands*
[3] *Nestlé Centre R&D Tours, Avenue Gustave Eiffel, 37097 Tours, France*
[4] *Zurich University for Applied Sciences, Department Life Sciences and Facility Management, Institute for Chemistry & Biological Chemistry, 8820 Wädenswil, Switzerland*
[5] *Nestlé Product Technology Center, 1350 Orbe, Switzerland*

Abstract

Considerable work has been devoted in the last decades to the identification and quantification of key aroma-active compounds in coffee as well as their precursors. The aim of this work was to demonstrate the applicability of a data-driven holistic method rather than a targeted chemical study. As an illustrative example, coffees at different roast degrees were analysed with a range of instrumental techniques (LC-MS, GC-MS, PTR-MS) and evaluated by a sensory panel. This allowed identifying correlations between chemical markers and sensory qualities and developing a deeper understanding on reaction mechanisms involved in coffee aroma formation.

Introduction

Already in the early 1970s, chemometrics led to the development of statistical methods to treat multivariate data sets obtained by chemical analysis (1,2), in parallel to the design of optimized measurement strategies. Most of the theories developed at that time are still used when dealing with multiple and multivariate datasets, even though today's computers allow the treatment of much larger volumes of data. The application of "omics" approaches to monitor metabolites in the human body related to various diseases accelerated the development of statistical and instrumental techniques. Minimalistic approaches, such as principal component analysis (PCA) and partial last squares (PLS) and their extensions to orthogonal-PLS (OPLS), hierarchical PCA, PLS and OPLS, with the aim to reduce a multidimensional space to a lower dimensional planes, are regularly used to investigate complex problems. A main advantage of "data driven" methods is that they are not based on fundamental chemical theories and can therefore be applied to reproducible unbiased data.

The application of chemometrics to coffee is interesting because of its complexity, e.g. the formation of coffee flavour during roasting, but also due to the success of chemometrics in linking quality differences to aroma compounds and precursors. The range of datasets that can be included in such studies is large and may encompass genetic fingerprints, agricultural information, meteorological data during bean maturation, chemical fingerprints and sensory profiles. Some of these data can directly be compared between samples (e.g. the number of days of sunshine or the growing region) while others need to be pre-processed. In particular, GC and LC data need to be pre-processed in such a way that peaks are recognized

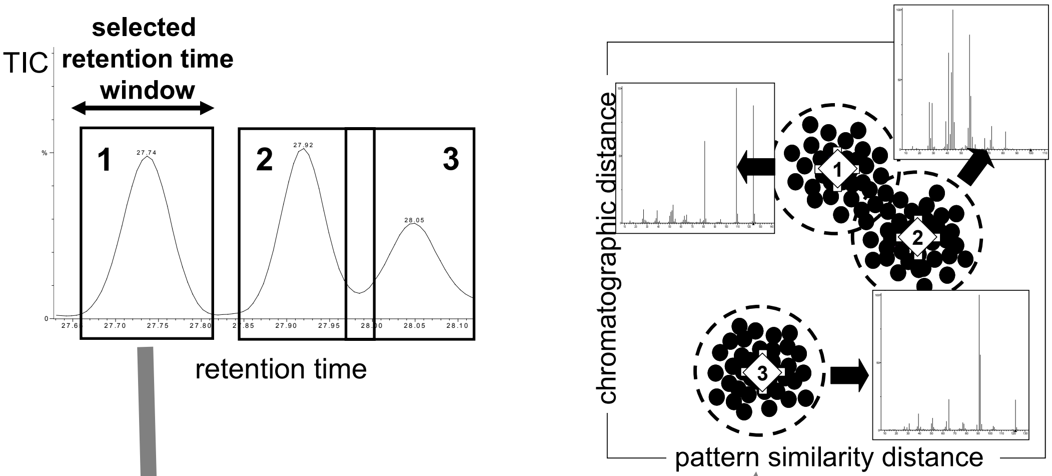
and aligned to compensate for shifts in retention time. A fully targeted approach requires that all compounds be identified before applying multivariate analysis. An untargeted approach overcomes these limitations but requires a more sophisticated pre-processing of the data including baseline correction, peak picking, alignment and centrotyping of raw data sets. As a consequence, time consuming identification can be focused on characteristic markers selected by multivariate statistics.

Experimental

To relate differences in chemical composition to cup quality, 65 coffee varieties grown in well defined conditions were evaluated by ten trained coffee panellists. Chemical data of volatile compounds was obtained by GC-TOF-MS/PTR-MS (Tenax desorption) and online PTR-MS measurements of roast and ground (R&G) coffee, prepared with an espresso machine. Volatiles released from coffee extracts within a sampling cell were analysed by online PTR-MS and trapped during two minutes on a Tenax trap for desorption on column using an automatic thermal desorption unit (*4*). Online PTR-MS data were interpreted by combining the GC-PTR-MS datasets with GC-MS (3). Thus, the various molecular contributions to single PTR-MS ion signals can be quantified and traced over time. Non-volatile compounds of extracts of green, slightly roasted and dark roasted coffees were analysed with LC-MS and GC-TOF-MS (after MSTFA derivatization) (6). For LC-MS measurements, 20 mg of powder from beans were weighed in 10 ml glass tubes and dissolved in 3 ml pure water and 75% methanol (containing 0.1% formic acid), respectively, by vortexing and sonication for 15 min. After centrifugation, the extracts were filtered through 0.2 µm PTFE filters, and directly used for LC-PDA-QTOF-MS analyses in ESI positive mode (6). For GC-TOF-MS measurements, 20 mg were weighed and extracted with pure water (80°C) in addition of an internal standard (Ribitol; Sigma, cat. no. 488-81-3). After stirring (10 min at 70°C in a thermomixer at 950 r.p.m.) the sample was centrifuged (10 min at 11000 g) and 750 µL chloroform (-20°C) added to the supernatant. The upper clear water phase (15 µL) was taken, dried in a vacuum concentrator and used for on-line MSTFA derivatization and GC-TOF MS analysis.

LC-MS and GC-TOF-MS raw data were processed by using the Metalign software *(www.metalign.nl)*. The software includes base line correction, peak picking respecting a limitation in signal to noise ratio and alignment of the detected peaks through all samples by an algorithm which compensates for slight shifts in retention time. Due to the ionization induced fragmentation by electron impact in GC-MS, single compounds are represented by an average of 10 mass signals. To reduce the data volume and eliminate redundant information a “centrotyping” program, similar to that reported in Ref. (5), was applied. This program correlates the intensity profiles of individual mass signals across all samples within a predefined retention time window which can be adjusted according to the shifts in retention time caused by the limitation of instrumental accuracy. Mass signals that correlate are clustered and expressed as single centrotype since they are expected to belong to one and the same compound (Figure 1). This reduces the data volume without loss of information, since the fragmentation pattern of each centrotype is stored. Hence, this information can be used for identification via comparison with databases.

Cluster analysis and correlation maps were obtained to visualize correlations between sensory data, volatile compounds and non volatile compounds. Correlation maps help identifying groups of related centrotypes and show their interrelation.



Peak Nr	Scan Nr	Ret(umin)	Mass	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	....	Sample 64	Sample 65
146	3235	10692450	33	437	473	513	434	506	529	....	301	330
2565	3235	10692450	56	58409	60879	58508	56705	57229	61151	....	36131	38724
2930	3235	10692450	58	3893	3911	3947	3826	3868	3920	....	2591	2630
2763	3236	10695730	57	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	45007	48893
3010	3237	10699000	59	2570	2542	2640	2450	2431	2688	....	1665	1785
3617	3237	10699000	65	317	335	327	324	325	342	....	240	224
193	3239	10705570	37	1050	1075	1088	1119	1097	1161	....	710	766
1785	3239	10705570	52	663	760	709	753	672	794	....	465	474
6357	3239	10705570	83	125	125	125	127	[illegible]	[illegible]	....	125	229
10	3240	10708850	29	1739	1932	1683	1220	1892	1795	....	977	1728
26	3240	10708850	30	[illegible]	1254	[illegible]	[illegible]	1305	1305	....	776	862
609	3240	10708850	41	89904	92971	90090	90183	[illegible]	[illegible]	....	59830	64841
1395	3240	10708850	49	218	175	[illegible]	150	141	200	....	146	[illegible]
1443	3240	10708850	50	2056	[illegible]	[illegible]	2240	1990	2099	....	[illegible]	1505
3331	3240	10708850	62	227	[illegible]	217	209	231	211	....	[illegible]	[illegible]
4647	3240	10708850	71	8291	8332	8147	[illegible]	8073	8601	....	5587	[illegible]
70	3241	10712120	31	31007	[illegible]	31014	31241	30658	[illegible]	....	20861	22039
386	3241	10712120	39	31258	32460	[illegible]	32018	31113	32674	....	20807	22425
1577	3241	10712120	51	2552	2588	2458	2671	2582	2628	....	1720	[illegible]
2192	3241	10712120	54	2589	2600	2702	2776	2569	2567	....	[illegible]	1769
3186	3241	10712120	61	422	403	382	461	477	453	....	125	125
4438	3241	10712120	70	119448	123905	118938	117944	117860	123359	....	79487	84488
255	3242	10715400	38	2756	3001	3013	2918	2755	2774	....	2021	2057
512	3242	10715400	40	5012	5230	5148	5067	5272	5345	....	3524	3567
769	3242	10715400	42	85578	87409	85238	84508	84316	88879	....	56159	60851
942	3242	10715400	43	70150	72318	70397	69630	69102	72398	....	45856	50339
1087	3242	10715400	44	5880	6295	5953	5970	5898	6125	....	3842	4266
1164	3242	10715400	45	19320	19555	19357	19277	18922	19909	....	12883	13801
1264	3242	10715400	46	7366	7615	7549	7234	7372	7744	....	5035	5181
1330	3242	10715400	47	1566	1602	1437	1504	1484	1518	....	1003	1114

Figure 1. *Example of eluted peaks selected within a defined retention time window. Three mass peaks correlate strongly through all samples and therefore most probably belong to the same compound. A fourth mass peak intensity shows a different pattern and therefore belongs to another compound.*

Results and Discussion

A large set of markers were aligned through all samples applying the Metalign program to GC and LC data. Further reduction of the data using a centrotype program helped remove redundant information, improving the readability and accuracy of correlation maps and multivariate analysis. This approach allowed comparing the differences in concentration of more than 200 volatile and 500 non-volatile compounds. The fragmentation patterns of individual compounds allowed identifying more than 150 volatile compounds. The identification of non-volatile compounds was focused on key correlations, where cluster analysis and correlation maps showed to be useful when investigating correlations between volatile

compounds their precursors and sensory profiles. The obtained datasets allowed developing a sensory predictive model which goes beyond the one already published (4), mainly by increasing the number of compounds included in the model (data not shown).

While the approach works for most of the compounds, volatile pyrazines were challenging to be differentiated by the automatic pre-processing and needed frequent manual intervention to avoid misalignment. However, the network of pyrazines was highly correlated (Figure 2) when analyzing the corresponding GC-TOF-MS correlation map.

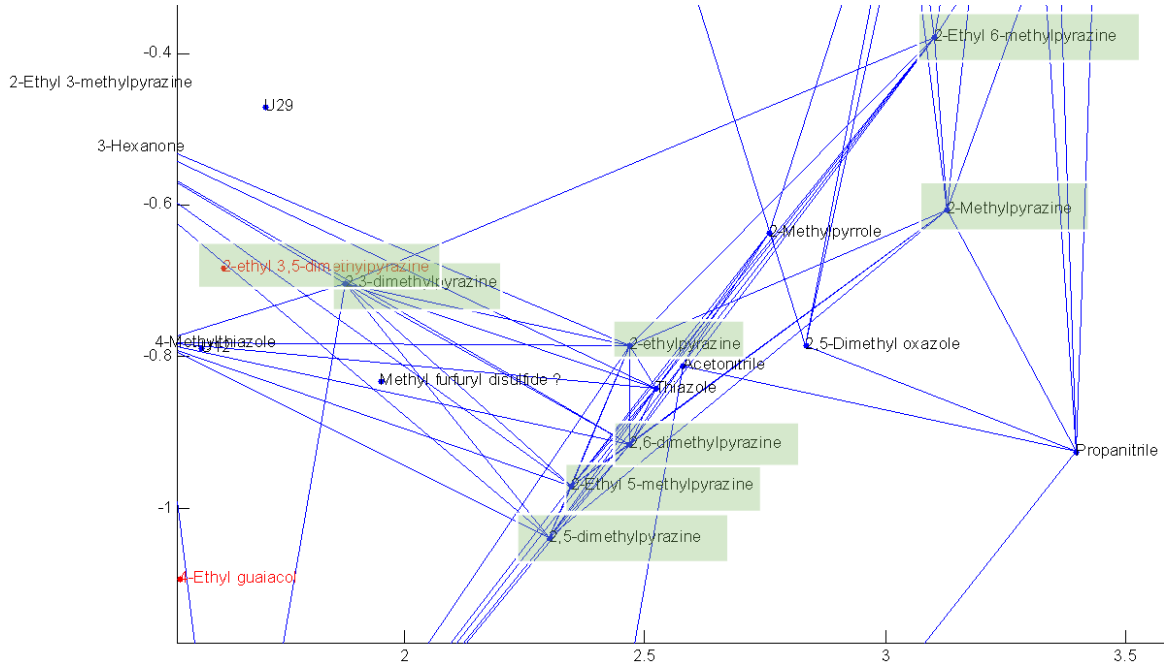


Figure 2. *Correlation map: This network shows a series of pyrazines that are highly correlated (blue lines indicate a correlation higher 0.9).*

By investigating changes in the chemical composition of volatiles and non-volatiles and testing their impact on predicted or evaluated sensory profiles, a robust method was developed to identify coffee varieties with the highest potential of in-cup quality.

References

1. Sharaf M.A., Illman D.L., Kowalski B.R. (1986) Chemometrics, Chemical Analysis Series Vol. 82: Wiley, New York.
2. Otto M. (2007) Chemometrics, Statistics and Computer Application in Analytical Chemistry: Wiley, Weinheim.
3. Lindinger C., Pollien P., Ali S., Yeretzian C., Blank I., Mark T. (2005) *Anal. Chem.* 77: 4117-4124.
4. Lindinger C., Labbe D., Pollien P., Rytz A., Juillerat M.A., Yeretzian C., Blank I. (2008) *Anal. Chem.* 80: 1574-1581.
5. Tikunov Y., Lommen A., de Vos R.C.H., Verhoeven H.A., Bino R.J., Hall R.D., Bovy A.G. (2005) *Plant Physiology* 139: 1125-1137.
6. de Vos R.C.H., Moco S., Lommen A., Keurentjes J.J.B., Bino R.J., Hall, R.D. (2007) *Nature Protocols* 2(4), 791.