Jan-Piet Franke,[1] *Ph.D.; Rokuas A. de Zeeuw,*[1] *Ph.D.; and Paul G. A. M. Schepers,*[1] *Drs.*

Retrieval of Analytical Data and Substance Identification in Systematic Toxicological Analysis by the Mean List Length Approach

REFERENCE: Franke, J. P., de Zeeuw, R. A., and Schepers, P. G. A. M., **"Retrieval of Analytical Data and Substance Identification in Systematic Toxicological Analysis by the Mean List Length Approach,"** *Journal of Forensic Sciences*, JFSCA, Vol. 30, No. 4, Oct. 1985, pp. 1074–1081.

ABSTRACT: Chromatographic techniques are basic tools in systematic toxicological analysis. Extensive data bases with retention parameters of known drugs to aid in the identification of substances found are available or in preparation. For a search in such a data base the computer is indispenable. The commonly used window search has some disadvantages which can be overcome by a search based on the statistical concept, the mean list length. The latter retrieval system gives for each candidate in the identification process a probability value. It is shown that these probability values are highly influenced by the reproducibility of the retention parameters of the analytical systems used. Explanations for these phenomena are given.

KEYWORDS: toxicology, drug identification, computers, chromatographic analysis, systematic toxicological analysis, substance identification, data retrieval, thin-layer chromatography, gas liquid chromatography

The basic aim of systematic toxicological analysis (STA) is to detect and identify substances of toxicological relevance in general unknown cases as well as in cases where the presence or absence of (a) given substance(s) has not been established beyond doubt. Chromatographic techniques, such as thin-layer chromatography (TLC), gas chromatography (GC), and high performance liquid chromatography (HPLC) are important tools in these analyses. For each of the chromatographic methods there are numerous systems available. For the selection of systems best suited for STA a statistical evaluation method was developed, the mean list length (MLL) concept [*1,2*]. This method is capable of computing an objective criterium to establish the identification power of a single system, a combination of systems, as well as combinations of different techniques, for instance combinations of TLC and GC [*2*]. Once the best systems and combinations of systems have been established it becomes meaningful to compile reference data for toxicologically relevant compounds in these systems and store them in a data bank so that they can be used for substance identification in unknown cases. Various data bases for such purposes have recently become available [*3–5*], whereas additional ones are in preparation.

Reliable identification of a substance can only be achieved if more than one analytical system is used. When searching in a large data base, using combinations of data observed in more

Received for publication 15 Oct. 1984; revised manuscript received 26 Feb. 1985; accepted for publication 3 March 1985.

[1]Senior research associate, professor of toxicology, and research associate, respectively, Department of Toxicology, State University, Groningen, The Netherlands.

than one analytical system, the computer is an indispensable tool. The retrieval method most often applied is the "window approach." Here it is determined for each substance in the data base if its value lies within a certain window around the value found for the unknown compound(s). If so, the substance is a possible candidate and the computer prints out a list of all possible candidates. This may have the disadvantage that no information is obtained on the closeness of the "match," that is, whether the candidate lies more towards the center or more towards the margin of the window. Moreover, substances just outside the window in one system are rejected and no longer considered in other systems, for example, when its concentration is below the detection limit.

The mean list length approach, which was initially developed to establish the identification power of a system, can also be used as a retrieval method, which overcomes the above disadvantages of the window search [*2*]. In this article the principle of the MLL approach for data retrieval and substance identification will be outlined and the main parameter influencing the outcomes of the search will be discussed.

Computer Retrieval System

The computer program for the data retrieval based on the MLL approach was written in Pascal for a Control Data Corporation Cyber 170/760 computer. A copy of this program is available upon request.

The Concept of Mean List Length

In the concept of MLL the assumption is made that the data obtained from the analytical procedure for a given substance show fluctuations, from day to day, and from laboratory to laboratory, which follow a certain, known distribution pattern. For instance, it is assumed that for TLC the corrected, standardized, R_f value of a given substance varies, following a normal distribution with the listed value (in the data base) as the mean value and with a given standard deviation, which varies from system to system. Now, when an unknown spot is found in the analysis, the corrected R_f value (R_f^c value) is derived and the probability of identification can be calculated for all substances in the data base. The further the retention value of a substance is the lower the probability will become. After normalization in such a way that the sum of the probabilities becomes 1, the substances can be ranked in decreasing order of probability. Then the top of the list will give substances with the highest probability. Table 1 shows an example for a spot with $R_f^c = 4$ in a TLC system with pure methanol as the solvent. Here only the first 13 candidate substances are shown but in principle the computer can print out probabilities of identification for all substances in the data base.

Obviously, a decision has to be made where to cut off the list. There are at least three ways to achieve such a decision:

- One can exclude all substances with a probability below a certain level, for instance 5%. Then, strychnine is the last substance of the list (Table 1, Column 2).
- The end of the list is reached when the cumulative probability exceeds a certain level, for instance 95%. For that level methylamphetamine is the last substance (Table 1, Column 3).
- On the basis of mathematical decision theory, the loss function criterium was developed [*1*]. All substances with a loss function less than 1 will appear on the list (Table 1, Column 4).

In these ways lists of candidates for identification can be established together with a probability value for each substance.

Similar calculations can be made for other techniques such as GC, where also a normal distribution of the Kovats retention index is assumed with a standard deviation of 10 to 20

TABLE 1—*Partial list of possible candidates for the identification of a spot in TLC with a* R_f^c *value of 4.*[a]

Substance	R_f^c	Probability, P, %	Cumulative Probability, F, %	Loss Function Criterium[b]
1. Naphazoline	4	15.3	15.3	0.01
2. Ametazole	5	14.1	29.3	0.02
3. Protriptyline	5	14.1	43.4	0.03
4. Atropine	6	11.1	54.5	0.05
5. Cyclopentamine	6	11.1	65.6	0.06
6. Antazoline	7	7.4	73.0	0.10
7. Desipramine	7	7.4	80.4	0.12
8. Strychnine	7	7.4	87.8	0.13
9. Guanethidine	0	4.3	92.1	0.23
10. Ephedrine	9	2.1	94.2	0.49
11. Methylamphetamine	9	2.1	96.3	0.51
12. Nortriptyline	9	2.1	98.4	0.53
13. Chlorpheniramine	10	0.9	99.3	1.25

[a]TLC: solvent methanol; silicagel plate. M: number of substances in the set under investigation.
[b]Loss function criterium: $F_i \cdot P_i^{-1} (M + 1 - i)^{-1}$.

retention index units. For a combination of systems the probability values per substance for each system can be multiplied, followed by normalization and ranking of the obtained values.

In this paper the length of the list with candidates is established with a combination of the loss function criterium and probability value at a cutoff level of 1%.

Discussion

To show some of the potentials and limitations of the computerized data search by the MLL approach a simple case is presented as an example. In a street drug sample cocaine and procaine were ultimately found to be present. For the initial screening TLC and GC were performed giving the results as outlined in Table 2. With the two peaks in GC and the two spots in TLC, two combinations of spots with peaks are possible: two configurations. For each of these configurations a search in the data base has to be executed. Initially, a search was performed in a small subset of the data base with 100 basic and some neutral drugs. Table 3 shows the results after a "window search" and Table 4 shows for Configuration 2 the results with different values

TABLE 2—*Analysis of a street drug sample.*[a]

TLC	GC
R_f values Standards: 24 70 85 Sample: 30 42	Retention indices Sample: 2010 2195
R_f values corrected/library Standards: 22 67 87 Sample: 27.9 39.6	

Configuration 1			Configuration 2		
TLC		GC	TLC		GC
27.9	/	2010	27.9	/	2195
39.6	/	2195	39.6	/	2010

[a]TLC: methanol-butanol (60:40), 0.1*M* NaBr on silicagel plates. GC: packed columns; 3% OV-1.

TABLE 3—*Results of a window search.*[a]

Configuration 1	Configuration 2
27.9/2010	27.9/2195
chlorpheniramine	atropine
methapyrilene	cocaine
thenyldiamine	mepyramine
tripelennamine	thonzylamine
39.6/2195	39.6/2010
amethocaine	dimethoxanate
carbetapentane	procaine
cocaine	thenyldiamine
doxepin	tripelennamine
imipramine	
mepyramine	
perphenazine	
phenindamine	
thonzylamine	

[a]Data from Table 2; 100 substances under investigation; window GC = ±40; window TLC = ±10.

TABLE 4—*Probabilities of identification for a MLL search for different values of SD(TLC).*[a]

Configuration 2			SD(TLC) 5	2.5	1
SAMPLE: 27.9/2195					
Candidates:	28/2199	atropine	33.7%	46.8%	90.6%
	30/2187	cocaine	28.5	30.2	8.8
	31/2203	thonzylamine	25.6	19.9	...
	33/2220	mepyramine	9.1	...	...
	40/2207	perphenazine	1.5	...	...
SAMPLE: 39.6/2010					
Candidates:	42/2018	procaine	35.6%	40.6%	23.1%
	36/1999	thenyldiamine	28.6	21.1	...
	38/2029	dimethoxanate	26.1	36.0	76.1
	34/1980	tripelennamine	7.5	...	...
	52/2020	cyclizine	1.8	...	...

[a]SD(GC) = 20; 100 substances in the set under investigation.

of the standard deviation of the TLC system using the MLL search. The windows for GC and TLC in Table 3 correspond to a standard deviation of 5 for TLC and 20 for GC. The values in the data bank are from data published in the literature and are obtained on an interlaboratory basis. Therefore, the chosen search windows are large but comparable with those recommended [*3,4*]. In Configuration 2 the first four substances for both sets were found with both search methods. With the window search perphenazine was not found because its R_f value lies just outside the window (and is rejected). Because of its retention index, which is close to the value found, it remains a candidate in the MLL approach. The same holds for cyclizine.

Obviously, when narrowing the search window or reducing the standard deviation of the system with both approaches, the number of candidates will diminish. With the MLL approach a reduction in the standard deviation results in a change in probabilities of the substances. Figure 1 shows a probability surface for the three main candidates of the second set in the second

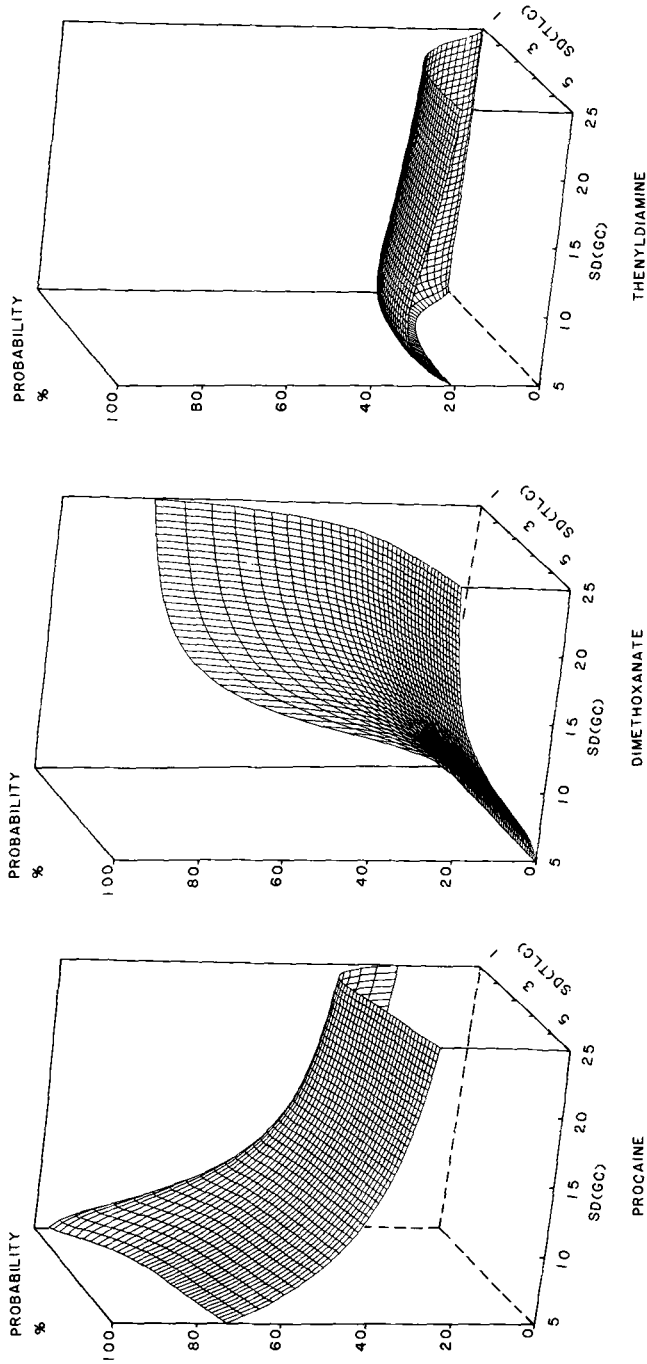

FIG. 1—*Combined influence of standard deviations of TLC and GC on the probabilities of identification for three substances based on experimental data in Configuration 2, Table 2.*

configuration. With a standard deviation (SD)(GC) = 25, an increase in SD(TLC) results for procaine and thenyldiamine first in an increase in probability, then followed by a slight decrease. For dimethoxanate only a decrease in probability is observed. To explain this phenomenon an overview of some data is given in Table 5. On the basis of the retention index alone, procaine will get the highest probability (the smallest difference with the experimental value) but, for the TLC data, dimethoxanate is the best candidate. This can also be expressed in terms of eccentricities (u). The u value is the number of standard deviation units between the listed value in the data base and the observed value. For SD(TLC) = 5 and SD(GC) = 20 these u values are comparable in magnitude. However, a decrease of SD(TLC) to 1 results in a rapid increase in the u values. Figure 2 shows the right halves of the probability distribution for SD(TLC) = 1, 2.5, and 5, respectively. Noted that the area under each curve is the same. For SD(TLC) = 1 the probability of dimethoxanate is much higher than for procaine, with the probability of thenyldiamine being practically zero.

When SD(TLC) is 2.5, the probability of dimethoxanate is slightly increased, but the probability of procaine is increased to a much larger extent and the same holds for thenyldiamine. However, relative to each other, the probability of dimethoxanate is decreased as a result of the

TABLE 5—*Eccentricities for different standard deviations of the TLC system.*

			u(GC)	u(TLC)		
	ΔR_l	ΔR_f	SD 20	SD 5	SD 2.5	SD 1
Procaine	8	2.4	0.4	0.5	1.0	2.4
Thenyldiamine	11	3.6	0.6	0.7	1.4	3.6
Dimethoxanate	19	1.6	1.0	0.3	0.6	1.6

[a]Experimental data: $R_l = 2010$; $R_f^c = 39.6$, see Table 2.

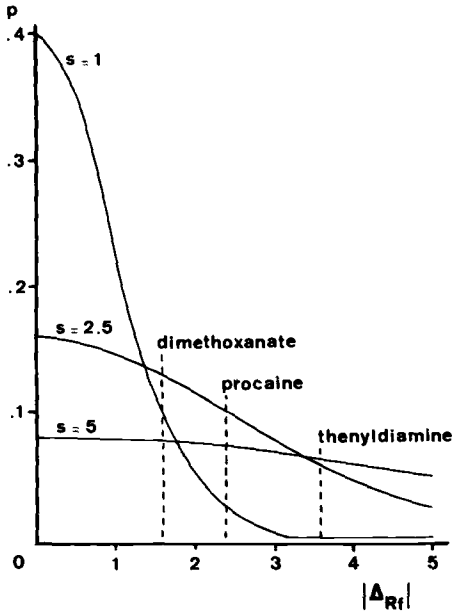

FIG. 2—*Probability distributions for three different values of standard deviations of the TLC system.*

large increase in probability of procaine and thenyldiamine. Increasing the SD(TLC) to 5 results in a decrease of the probabilities of procaine and dimethoxanate whereas the value of thenyldiamine remains about the same. Relative to each other, a small increase of probability of thenyldiamine can be seen (Fig. 1).

Figure 1 shows also that the curvature of the probability surface for all these substances is most pronounced at low SD values for both GC and TLC. In this region, a small change in SD will have a great impact on the probability values of the different substances. For instance, with SD(GC) = 5 and SD(TLC) = 1 the probability of procaine is over 95% (Fig. 1) whereas the probabilities of the other substances are very low (<5%). Thus, one would conclude one candidate present (procaine). With SD(GC) = 10 and SD(TLC) = 2 the probability of procaine drops to about 60% and now thenyldiamine and dimethoxanate are both candidates with about 20% probability each. Therefore, in cases with low SD values, the probability outcomes must be judged carefully.

On the other hand, too high values of SD will increase the number of candidates beyond necessity, so that much more work has to be done to establish the identity of the underlying compound. Thus, a proper determination of the reproducibility of the chromatographic technique is of great importance for data retrieval.

Conclusion

For data retrieval and substance identification in STA by means of an extensive data bank the computer is an indispenable tool and the MLL approach will give more information than just the names of the substances. As the reproducibilities of the analytical systems used play an important role in the outcomes of the search, more work has to be done to establish intra- and inter-laboratory variability of the different analytical systems. Obviously, the outcome of a search in a data base is highly dependent on the quality of the data base or on the quality of the laboratory performance in producing the retention data or both. With conventional search methods using the "window" approach, it is hardly possible to detect the compound involved, if in the data base for a particular analytical method an outlying result is incorporated or if the searching laboratory has produced an outlying result itself. Enlarging the search window to find such a substance is not an acceptable solution, since then too many irrelevant substances will be included.

However, a search with the MLL approach, using a relatively small standard deviation will be able to find a substance with an outlying value in a system as long as more than one system is being used for the search. Therefore, it is important to pay attention also to substances with relatively low probabilities of identification. Further, confirmation of the identity can then be obtained by applying one or more additional analytical techniques.

In the example presented here only two compounds were present, so that with two analytical systems only two combinations of spots with peaks had to be taken into account. It must be borne in mind, however, that with an increasing number of spots or peaks or both, which is not uncommon in modern toxicology, the number of configurations increases exponentially. With 5 substances present and using 2 systems there are 120 different configurations. Application of a third analytical system to such a case will increase the number of configurations to more than 14 000.

References

[1] Akkerboom, J. C., Schepers, P., and Werff, J. van de, "Thin Layer Chromatography, a Case Study," *Statistica Neerlandica*, Vol. 34, No. 4, 1980, pp. 173-187.

[2] Schepers, P., Franke, J. P., and Zeeuw, R. A. de, "System Evaluation and Substance Identification in Systematic Toxicological Analysis by the Mean List Length Approach," *Journal of Analytical Toxicology*, Vol. 7, No. 6, Nov./Dec. 1983, pp. 272-278.

[3] Finkle, B. S., Franke, J. P., Moffat, A. C., Moeller, M., Mueller, R. K., and Zeeuw, R. A. de, "Gaschromatografische Retentionsindices toxikologisch relevanter Verbindungen auf SE-30 oder OV-1," Mitteilung I der Kommission für klinisch-toxikologische Analytik der Deutsche Forschungsgemeinschaft, Verlag Chemie, Weinheim, Deerfield Beach, FL, Basel, 1982.

[4] Stead, A. H., Gill, R., Wright, T., Gibbs, J. P., and Moffat, A. C., "Standardized Thin Layer Chromatographic Systems for the Identification of Drugs and Poisons a Review," *Analyst*, Vol. 107, No. 1279, 1982, pp. 1106–1168.

[5] Ardrey, R. E. and Moffat, A. C., "Gas Liquid Chromatographic Retention Indices of 1318 Substances of Toxicological Interest on SE-30 or OV-1 Stationary Phase," *Journal Chromatography*, Vol. 220, No. 3, 1981; *Chromatographic Review*, Vol. 25, No. 3, pp. 195–252.

Address requests for reprints or additional information to
J. P. Franke, Ph.D.
Department of Toxicology
State University
Antonius Deusinglaan 2
9713 AW Groningen, The Netherlands

Erratum

In the Oct. 1985 *Journal*, on p. 1074, Dr. de Zeeuw's first name was inadvertently misspelled as Rokuas. The correct name is Rokus. We are sorry for the inconvenience this has caused.