Turk J Chem
26 (2002) , 45 – 58.
© TÜBİTAK

Automated Recognition of Ergogenic Aids Using Soft Independent Modeling of Class Analogy (SIMCA)*

M. PRAISLER
Department of Physics, University of Galati,
Domneasca St. 47, 6200 Galati, Romania
e-mail: studium@sisnet.ro
J. Van BOCXLAER, A. De LEENHEER
Laboratory of Medical Biochemistry and Clinical Analysis, Ghent University,
Harelbekestraat 72, B-9000 Gent-BELGIUM
D. L. MASSART
Laboratory of Pharmaceutical and Biomedical Analysis, V rije Universiteit Brussel,
Laarbeeklaan 103, B-1090 Brussel-BELGIUM

Received 25.10.2000

The introduction of effective drug testing procedures in doping control reduced, but did not eliminate, their abuse by athletes. The most important analytical challenge is the recognition of the new analog compounds. In attempts to circumvent existing controlled substance laws, slightly modified chemical structures (by adding or changing substituents at various positions on the banned molecules) are used. As a result, no substance belonging to a prohibited class may be used nowadays, even if it has not been specifically listed. We present a chemometric procedure acting as an automated GC-FTIR screening test detecting the molecular structural similarity of unknown compounds with the main classes of ergogenic aids. The knowledge base defining the reference GC-FTIR spectral patterns has been built according to criteria encompassing toxicological, pharmacological and neurochemical aspects. The class identity of a compound is diagnosed within seconds, using Soft Independent Modeling of Class Analogy (SIMCA). The predictive value of the system was assessed at a testing accuracy of 95%. Compounds giving cross-reactions with traditional screening techniques produce a negative result. The specificity and the selectivity of the screening test, evaluated by testing 160 toxicologically relevant compounds, are discussed, emphasizing the chemical and physical factors affecting these parameters. The specificity of the system recommends the procedure as a highly specific, selective, fast, and user-friendly screening test, which screens for ergogenic aids found in powders, tablets or solutions with a prediction accuracy adequate for investigations in analytical toxicology and doping control.

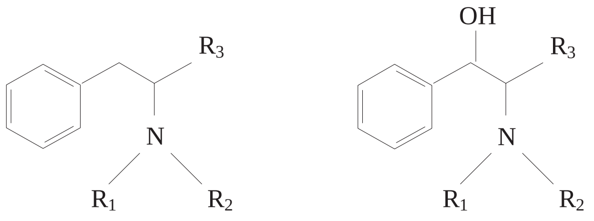
Key Words: Principal component analysis, Soft Independent Modeling of Class Analogy, Infrared spectrum, Amphetamines, Drugs of abuse

*This paper has beed presented at MBCAC III (3rd Mediterranean Basin Conference on Analytical Chemistry) 4-9 June, 2000 Antalya-Turkey

Introduction

Doping in sport does not only threaten to damage sport as a social institution, but it is also detrimental to the fundamental ethical values, namely, fair play, integrity and solidarity. The protection of the health of the athlete is also of ethical concern, for it is endangered by an enormous pressure on the athlete to push towards an even higher level of performance. Doping in sport touches on medical ethics, as sport doctors today are called upon to help enhance sports performance by offering medical substances and specific methods for more than simply therapeutic reasons, and in a way that is not readily transparent.

Athletes are seeking a competitive advantage by using various substances that have been dubbed "ergogenic aids". Doping is considered to have occurred when substances belonging to prohibited classes of pharmacological agents are administered. Amphetamines, sympathomimetic amines and cocaine (see Figure 1) represent the main classes of abused substances. These stimulants constitute more than 85% of positive cases reported by the International Olympic Committee (IOC) Medical Commission statistics. According to the latest reports of the European Group of Ethics, an advisor to the European Commission, better methods for detecting the different types of doping practiced by both amateurs and professionals are seen as crucial for the future of international competitive sports[1].



Amphetamine analoges

Ephedrine analoges

Figure 1. Molecular structures of the main stimulants abused by athletes.

Psychomotor stimulants exert their effects[2] primarily by potentiating actions on catecholamine cells and receptors in the central nervous system and in the periphery. The most commonly abused stimulants include amphetamine (β-phenylisopropylamine), methamphetamine (methyl-β-phenylisopropylamine), ephedrine (α-[1-(methylamino)ethyl]benzene-methanol) and cocaine ([1R-(exo,exo)]-3-(benzoyloxy)-8-methyl-8-azabicyclo[3.2.1]octane-2-carboxylic acid methyl esther), although there are many others whose abuse may depend on their pharmacological characteristics. Drugs of this class are used for a variety of therapeutic applications including weight loss, and the treatment of narcolepsy and attention deficit disorder with hyperactivity. Cocaine is used clinically as a local anesthetic. In general, psychomotor stimulants produce their pharmacological actions by enhancing dopaminenergic, noradrenergic and adrenergic transmission[3]. The result is increased circulating and synaptic concentrations of monoamines leading to overstimulation of central and peripheral receptors. The central actions of the psychomotor stimulants include increased awareness, decreased fatigue, and a feeling of well-being or euphoria. Their autonomic effects include tachycardia, pressor effects and vasoconstriction. In addition, they possess anorectic properties. Both amphetamine and cocaine can be toxic at higher doses. Acute toxic effects can include tremors, confusion, hallucinations, and diaphoresis. Lethal overdose can result from the cardiovascular effects of these drugs (e.g., cardiac arrest, cerebrovascular infarction or aortic rupture) or from respiratory depression, hyperpyrexia, and convulsions[4].

In attempts to circumvent existing controlled substance laws, clandestine laboratories are synthesizing slightly modified chemical structures by adding or changing substituents at various positions on the molecule of the parent compound (controlled substance). Most of the psychomotor stimulants are suitable for clandestine laboratory production and molecular synthetic modifications are easily accomplished, resulting in a number of homologs and analogs. Even the Internet now provides answers to anyone tenacious enough to search for a simple method to synthesize any analog or homolog of a phenethylamine. Many of the analogs of these stimulants have become controlled substances after evidence of their potential toxicological effects. As a result, the most recent trend in legislation is to define the term *controlled substance analog* as a chemical structure substantially similar to the structure of a controlled substance, which has a stimulant, depressant or hallucinogenic effect on the central nervous system that is substantially similar to or greater than the effect of a controlled substance[2]. With the introduction of drug testing programs, forensic laboratories have increasingly been involved with the analysis and identification of amphetamines and related analogs, paying special attention to novel structures.

As amphetamines are reasonably volatile substances, vapor-phase infrared spectroscopy has recently become, due to its specificity, one of the important instrumental methods for the identification of controlled substances in tablets and powders[5−7]. Different functional groups and molecular interactions brought about by symmetrical and asymmetrical molecular stretching vibrations and in-plane and out-of-plane bending vibrations result in a number of absorption bands. The infrared spectrum of a suspected drug results in a specific pattern that can be used to determine the identity of a compound, the spectral pattern being unique to the chemical structure of the drug. The absolute identification is performed through infrared fingerprinting, by comparing the pattern in the spectrum of an unknown with reference spectra of a primary drug standard. If the two spectra match within the limits of scientific certainty, identification is possible.

The combination of gas chromatography (GC) with Fourier transform infrared spectroscopy (FTIR) has enormously increased the resulting number of spectra. Library search systems are currently commercially available from most instrument manufacturers. Automated compound identification is usually done by calculating the Euclidean distance between the spectrum of the unknown and each spectrum present in the library. The identity of the library spectrum that is most similar to that of the unknown (has a maximal hit quality index HQI) is then assigned to the unknown. However, the Euclidean distance based IR library search system is unfortunately of little use when novel structures are investigated in confiscated drug samples. The search yields an incorrect answer when the unknown is not present in the library, although the closest "hit" resulting from the library search may be structurally similar to the unknown. This weakness becomes especially important in the case of vapor-phase FTIR libraries, as they are rather rare on the commercial market and less expanded in certain fields than the conventional condensed-phase FTIR libraries. The alternative is the in-house creation of field-oriented, but smaller, spectral libraries (such as the toxicology-oriented library of vapor-phase spectra created at the Laboratory of Toxicology, Ghent University, Belgium)[8].

Another major disadvantage of the Euclidian distance based identification system is that it does not provide any information if the HQI of the unknown not present in the library and the most similar compound found in the library exceeds the limits defined by the structural similarities of compounds belonging to the same class. Indeed, such information would allow the system to act, at least, as a (non-traditional) screening test. Assigning the class identity to a compound that cannot be found in the library for absolute identification is a very helpful analytical step, as it can streamline further determinations and thus save time, as well as significantly diminish the analysis costs. The assignment of the class identity of compounds has been solved in recent years by applying chemometric multivariate techniques with better performances than the traditional (Euclidean distance based) library search procedure[9,10]. Among these, principal component analysis (PCA) is a helpful tool when the data structure needs to be conserved[11]. As opposed to linear discriminant analysis (LDA) or to the partial-least-squares method (PLS), no information about categories or chemical substructure influences the PCA calculation.

The results of a PCA exploratory analysis also allow initiation of a multivariate classification. Although PCA score plots can be used for classification, discrimination as such can be better performed using specialized algorithms, such as Soft Independent Modeling of Class Analogy (SIMCA), which is a classification algorithm using disjoint class modeling by supervised pattern recognition[10]. The main goal of classification is to reliably assign new samples to classes that are characterized by PCA models. In addition, SIMCA can indicate the most discriminant variables, as well as the most important ones in defining the similarity among the members of a class of substances. The algorithm can also be reformulated as a multivariate outlier test, being a useful tool for classification refining purposes. Finally, SIMCA classification also has the advantage that it can decide whether unknown samples belong to more than two modeled classes.

In this study we present a new multivariate approach to automated class identity assignment, designed for the analysis of the most important ergogenic aids. It consists of a combination of PCA and SIMCA techniques, which allowed the development of a tailor-made, knowledge-based classification system for the class of amphetamines and the class of sympathomimetic amines. The two classes were modeled on the basis of structure-activity relationships. The procedure was designed with the intention of identifying novel analogs in drug samples not present in the spectral library.

Experimental

A Perkin Elmer (Buckinghamshire, U.K.) Autosystem GC was interfaced with a light pipe GC-IR System 2000 and connected to a FTIR System 2000 with a mid-infrared source and a medium band liquid nitrogen-cooled mercury cadmium telluride (MCT) detector. Temperature-programmed separations were carried out on a Hewlett-Packard (Palo Alto, CA, USA) Ultra-1 methylsilicone capillary column (25 m x 0.32-mm i.d., 0.52-μm film thickness). The carrier gas was helium at a flow rate of 1.8 ml/min. The analytical column outlet stretched into the light pipe inlet. Helium carrier gas was added as make-up gas at a flow rate of 1.8 ml/min at the connection between the capillary column and the light pipe. The gold-coated light pipe (12 cm × 1 mm i.d.) was heated at a constant temperature of 270°C.

Methanolic stock solutions (1.0 mg/ml) of the reference standards were injected into the GC-FTIR system. Chromatograms were calculated by the Gram-Schmidt vector orthogonalization method. Gram-Schmidt reconstruction was performed using 10 basis vectors throughout the run. Real time spectra were obtained by the addition of two scans, with a spectral resolution of 8 cm^{-1} and 32 background scans. The scan range was from 4000 to 580 cm^{-1}. Baseline correction was performed on the reconstructed Gram-Schmidt chromatogram (GS) and low-noise vapor-phase FTIR spectra were generated after co-addition.

The obtained reference vapor-phase FTIR spectra were stored in a computer-based library after normalization. The normalization procedure involves scaling each spectrum so that peak absorbance of the most intense band is set to unity. The spectral data were stored in the database at 5 cm^{-1} intervals. All spectra were reduced in size by eliminating the wavenumber intervals where the compounds in the database had no IR absorptions. Hence, the data used for multivariate analysis ranged from 3750 to 2550, and from

2000 to 600 cm^{-1}. The 160 samples and remaining 523 wavenumber intervals of 5 cm^{-1} resulted in a data matrix with 160 x 523 data.

The knowledge-based classification and identification system for the stimulants analogs was obtained as follows: a training set was selected from the database using criteria encompassing toxicological, pharmacological and neurochemical aspects, and a feature weight function w_k(I, II) was calculated on their basis; PCA was performed on the feature weighed spectra of the training set; clusters (subclasses) were identified and characterized using the principal components; a PCA model was calculated using the feature weighed spectra of each identified subclass of compounds; SIMCA classification was performed using a validation set formed with the feature weighed spectra of 160 toxicologically interesting compounds; and the quality of the classification was evaluated on the basis of the total correct classification rate (classification matrix) and of the percentage of compounds classified with a 5% significance level.

Results and Discussion

The initial training set consisted of two classes of compounds: class I contained the spectra of the first 10 amphetamine analogs (subclass code A) listed in Table 1 and the 6 ephedrine analogs (subclass code E) listed in Table 2. Class II contained 16 counterexamples (class code N) with very different molecular structures, selected with the intent of spanning a broad range of various toxicologically interesting compounds (bemegride, β-butyrolactone, cadaverine and its heptafluorobutyric (HFB)-derivative, codeine and its pentafluoropropionic (PFP)-derivative, caffeine, γ-butyrolactone, the trimethylsilyl (TMS)-derivative of γ-hydroxy butyric acid, the TMS-derivative of γ-hydroxy valeric acid, γ-valerolactone, nicotamide, piracetam, putrescine and dextromoramide, nicotine).

Table 1. Stimulant amphetamines of classification category A* with their spectral library entry code (ID), acronym, and substituents.

	ID	Name of the compound	Acronym	R_1	R_2	R_3
2-phenylethylamines						
	A7	amphetamine	AMP	H	H	$-CH_3$
	A12	benzphetamine		$-CH_3$	$-CH_2-C_6H_5$	$-CH_3$
	A17	β-phenylethylamine	BPEA	H	H	H
	A32	dextroamphetamine		H	H	$-CH_3$
	A46	phentermine		H	H	$-(CH_3)_2$
	A47	1-phenyl-2-butanamine		H	H	$-CH_2-CH_3$
	A74	methamphetamine	MAMP	H	$-CH_3$	$-CH_3$
	A96	N-ethylamphetamine	EAMP	H	$-CH_2-CH_3$	$-CH_3$
	A112	N,N-dimethyl-1-phenyl-2-ethanamine		$-CH_3$	$-CH_3$	H
	A114	N-n-propylamphetamine	PAMP	H	$-(CH_2)_2-CH_3$	$-CH_3$
	D45	fenproporex		H	$-(CH_2)_2-CN$	$-CH_3$
	D71	mephentermine		H	$-CH_3$	$-(CH_3)_2$
	D159	clobenzorex		H	$-CH_2-(C_6H_4Cl)$	$-CH_3$
1-phenylethylamines						
	D4	α-phenylethylamine	APEA	H	H	
	D102	N-methyl-α-phenylethylamine[a]	MAPEA	H	$-CH_3$	

Feature weight

In cases such as ours, when spectra were recorded using standard samples and the noise affecting data is very low, a feature weight spectrum can be useful. Feature weighing the infrared spectra may improve the separation of the two classes of compounds, leading to a higher prediction accuracy[12]. A feature weight spectrum w_k(I, II) was calculated (see Figure 2) for each wavenumber from the ratio of the intercategory variances to the sum of the intracategory variances:

$$w_k\left(I, II\right) = \frac{\sum A_I^2/N_I + \sum A_{II}^2/N_{II} - 2\sum A_I \sum A_{II}/N_I N_{II}}{\sum \left(A_I - \bar{A}_I\right)^2/N_I + \sum \left(A_{II} - \bar{A}_{II}\right)^2/N_{II}} \tag{1}$$

where A_I, A_{II} are the absorbances at wavenumber *k* for samples of classes I and II and N_I, N_{II} the numbers of samples of classes I and II. The greater the discrimination ability of a measurement at a particular wavenumber, the greater the feature weight; if a measurement has no discriminating power, w_k(I, II) = 1.

Table 2. Stimulant sympathomimetic amines (ephedrine analogs) of classification category E* with their spectral library entry code (ID), acronym, and substituents.

	ID	Name of the compound	Acronym	R_1	R_2	R_3
	E13	benzephedrine		$-CH_3$	$-CH_2-C_6H_5$	$-CH_3$
	E39	ephedrine		H	$-CH_3$	$-CH_3$
	E104	*N*-methylephedrine		$-CH_3$	$-CH_3$	$-CH_3$
	E110	norephedrine		H	H	$-CH_3$
	E111	norpseudoephedrine		H	H	$-CH_3$
	E129	pseudoephedrine		H	$-CH_3$	$-CH_3$

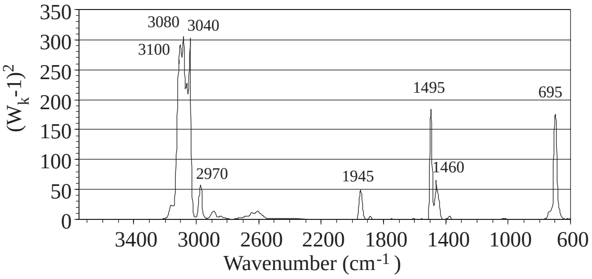


Figure 2. Optimum feature weight.

The main advantage of using an optimized discriminating feature weight is increased separation between the two initial classes of compounds (class I and class II). Which feature acts best depends on the nature of the compounds (the stability of the band parameters in their FTIR spectra) and on the number of samples (spectra) available. Three functions of feature weights reported to be successful in the literature[12,13] were investigated: w, w^2, $(w-1)^2$. PCA was performed on the original training set (no pretreatment), and on each scaled (feature weighed) training set. The best separation was found with $(w-1)^2$. An explanation might be the fact that, as opposed to the other functions, $(w-1)^2$ acts as a wavenumber selector by eliminating the signal ($(w-1)^2 = 0$) with no discrimination power ($w = w^2 = 1$). At these wavenumbers, the spectra from classes I and II present similarities negatively influencing their separation, or the spectra from the same class present important band parameter variations affecting the modeling of their similarity. By comparing the feature weight (w-1)^2 with the spectra of methamphetamine and ephedrine shown in Figure 3, we may observe that the weighing function enhances the absorptions that are the most similar in the spectra of the stimulants present in class I:

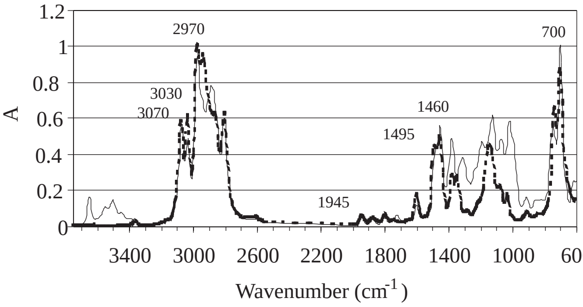


Figure 3. The vapor phase FTIR spectra of ephedrine (———) and of methamphetamine (- - - - - - - -).

- the absorption bands in the 3100 and 3000 cm^{-1}associated with the C-H stretching vibrations of the aromatic ring;

- the group of bands in the 3000-2750 cm^{-1} region associated with the stretching vibrations of C-H bonds in CH_2 or CH_3groups;

- the group of weak overtone and combination bands in the 2000 - 1900 cm^{-1} region specific to the aromatic ring, and which are characteristic of its type of substitution;

- the bands between 1535 and 1400 cm^{-1} associated with bands specific to the breathing vibrations of the aromatic ring, with N-H bending vibrations of the amino groups and with saturated C-H deformation vibrations;

- the bands associated with C-H out-of-plane bending vibrations below 800 cm^{-1} characteristic of the substitution pattern of aromatic compounds.

The form of the feature weight (w-1)^2 has two interesting aspects. First, the relatively important discrimination power found for these bands proves that the intensity of the absorption bands is much less important, for defining similarity within a class, than their specificity. Indeed, although so weak, the group of weak overtone and combination bands in the 2000 - 1900 cm^{-1} region show a remarkable intraclass stability of band parameters (see Figure 3) and are consequently enhanced by the feature weight as much as the strong 2970 cm^{-1} band. Secondly, the absence of any peak in the feature weight specific to the compounds in class II confirms the randomness (lack of similarity) of the molecular structures and associated spectra of these compounds. The same situation is encountered with the bands found between 3650 and 3500 cm^{-1} associated O-H stretching vibrations, which are present only in the spectra of the ephedrine analogs included in class I, and thus are not found to be specific to class I as a whole. In conclusion, feature weighing enhances the absorptions specific only to the common structural unit of the stimulants forming class I, i.e. the basic skeleton containing an aromatic ring linked by an aliphatic side chain to an amino group.

Exploratory analysis (PCA)

PCA was performed on the (w-1)^2 weighed spectra of the training set using the software package *The Unscrambler*® (Camo AS, Norway). Data were mean-centered to ensure that all results will be interpretable in terms of variation around the mean. The validation method was full cross-validation.

The way explained variances vary according to the number of model components was studied to decide how complex the PCA model should be. The explained variance, measured as a percentage of the total variance in the data, is a measurement of the proportion of variation in the data accounted for by the current PC. Usually only the first PCs contain genuine information, the later PCs being likely to describe mostly noise. A residual variance local minimum could not be reached for models with up to 20 PCs. Models built with more than two PCs did not significantly improve the explained variances of the first PCs, nor the final results (an effective discrimination of classes of interest and unknown compound classification). In such a case, a less complex model allows an easier interpretation of the PCA results, and ensures that noise has not been mistaken for information. We concluded that the exploratory analysis should be done with the first two PCs, which expressed 97% and 1% of the information respectively.

The influence plot (leverages vs. sums of X squared residuals) was used to detect outliers. As a result, benzylephedrine (E13), *N*-methylephedrine (E104), benzphetamine (M12) and β-phenylethylamine (M17) were discarded for the optimization of the discrimination among the three clusters observed in the score plot. The influence plot obtained with the remaining samples from the training set is illustrated in Figure 4. The compound with the largest leverage is dextroamphetamine (D32), the d-isomeric form of amphetamine. The large leverage is due to the fact that its absorption in the 3030-3045 cm^{-1} region (C-H stretching vibrations of the aromatic ring) is much stronger than the absorption in the spectra of the rest of the amphetamines that represent racemic mixtures (see Figure 5). The compound with the largest residual X-variance was phentermine (M46, α,α-dimethylphenethylamine), probably due to the specific absorptions generated by the two adjacent methyl groups linked to a carbon atom present in its structure as opposed to the other modeled amphetamines.

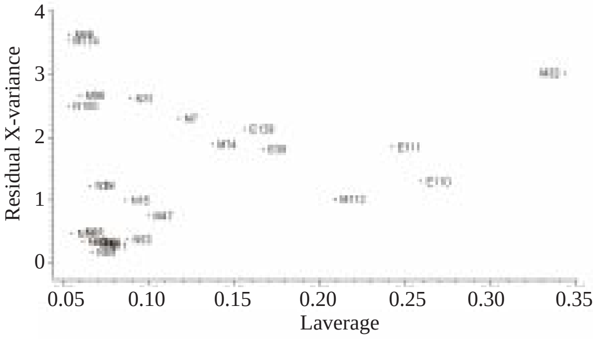


Figure 4. The influence plot used for the detection of outliers.

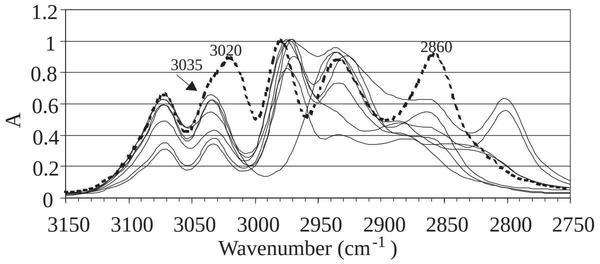


Figure 5. Atypical absorptions in the vapor phase FTIR spectrum of dextroamphetamine (.........), the d-stereoisomer of amphetamine, in comparison with the spectra of the amphetamine analogs representing d- and l-racemic mixtures (———).

The score (scatter) plot of the PCA analysis is shown in Figure 6. It shows that the amphetamines (subclass code A) are very well separated from the rest of the compounds, being the only ones characterized by positive PC1 scores and negative PC2 scores. Sympathomimetic amines (subclass code E) cluster the best, and are characterized by the high positive PC1 and PC2 scores. The nonstimulants (class code N) are best discriminated by negative PC1 scores. Interestingly, the nonamphetamines form a well-defined cluster, as if they were structurally very similar. In fact, their clustering is not due to a structural similarity such as the *presence* of a given structural unit, but to the common structural characteristic represented by the *absence* of the phenylethylamine skeleton.

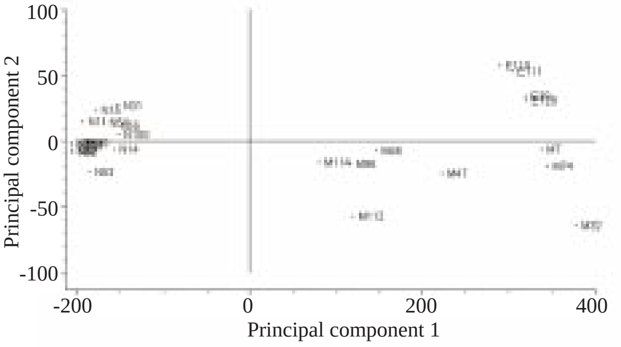


Figure 6. Score plot emphasizing the discrimination among amphetamine analogs (sample code A), ephedrine analogs (sample code E) and their counterparts (sample code N).

The loading of a variable on a PC reflects both how much the variable contributed to that PC, and how well that PC takes into account that variable's variation over the data points. Loadings also describe the relationship between variables. The loading plot obtained in our study is illustrated in Figure 7 and confirms that the variables with the highest discrimination power, as measured by their loadings, are those selected by the feature weight. The variables with the highest PC1 loadings (3030-3095 and 690-705 cm^{-1}) have the most important contribution to the discrimination of the A and E stimulants from the N compounds. The variables with the highest PC2 loadings (3030-3040 and 690-705 cm^{-1}) are the most important for the discrimination among stimulants, the discrimination being ensured by significant differences in the intensity of these absorptions displayed in the spectra of A and E stimulants respectively. An interesting fact is that in most of the cases, the correlated variables had consecutive values (such as 3065, 3070, 3075, 3080 and 3085 cm^{-1} for positive PC1 loadings; 690, 695, 700, 705 cm^{-1} for positive PC2 loadings; and 3030, 3035 and 3040 cm^{-1} for negative PC2 loadings). The size of these variable intervals is comparable with the widths of the corresponding absorption bands. This behavior suggests that the good discrimination between the three subclasses of compounds (A, E, N) was possible for two reasons. The system has taken into account not only the band wavenumbers where the maximum absorptions take place, but also the profiles of those IR bands, shown by the correlation among consecutive wavenumbers. In conclusion, additional spectra deresolving was not considered appropriate and all variables selected by the feature weight were kept in the data matrix. The two subclasses (A, E) of class I could be reliably put in evidence only by keeping all these variables into the data matrix. Indeed, simpler PCA models, based on a smaller number of variables, could be obtained only at the expense of the number of subclasses put in evidence and of the quality of the clusters' discrimination.

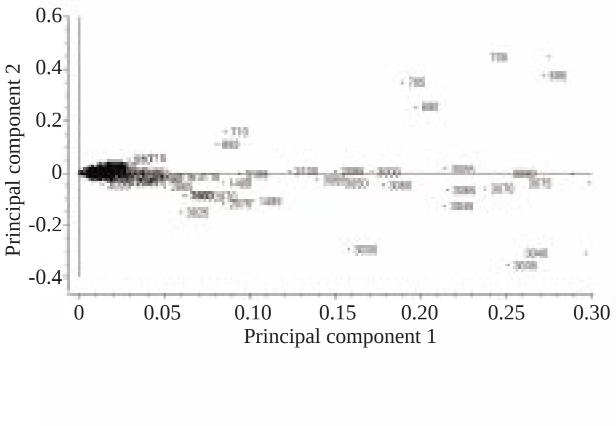


Figure 7. Loading plot identifying the variables with the most important discrimination power.

SIMCA classification

As the exploratory analysis proved the feasibility of discriminating among amphetamines and sympathomimetic amines, as well as their recognition from nonstimulants, a separate PCA model was computed for the amphetamines (model **A**), for the sympathomimetic amines (model **E**) and for their counterparts (model **N**). In each case, PCA was performed with the feature weighed spectra in each of the three subclasses defined within the training set. Data were mean-centered. Scaling based on the variables' standard deviation was performed in order to give all variables the same chance to influence the estimation of the components. The validation method was full cross-validation. A number of three PCs were calculated for models **A** and **N**, and two for model **E** as the number of E samples was smaller. A new (validation) set was composed with the 160 compounds in the laboratory-made spectral library. All spectra were recorded and processed (baseline corrected, normalized, and feature weighed) in the same way as those in the training set. The final matrix contained 160 x 523 data.

In the case of the PCA run for the exploratory analysis, when only one PCA model was calculated for all the compound subclasses in the training set, the best results were obtained when scaling was *not* applied after feature weighting (variable weight was taken equal to 1 so that both effects of $(w\text{-}1)^2$ could be present). On the other hand, SIMCA classification run with PCA models calculated for each subclass using unscaled data gave poor classification results. Much better results were obtained when the individual

PCA models were calculated with data scaled after feature weighting. Scaling counteracts the changes made by the feature weight in the absorptions measured at the selected discriminating wavenumbers. This indicates that while for the exploratory analysis (PCA) both effects of (w-1)^2 were needed (the feature weight enhancing the signal at the discriminating variables, and acting as a variable selector), in the case of SIMCA classification only the second effect (variable selection) was beneficial. Stressing the intraclass differences in absorption intensity caused by the feature weight influenced negatively the modeling power of the selected discriminating variables for the given subclass of compounds. Thus, in the case of the PCA models computed for individual subclasses, the first effect of the feature weight had to be balanced by scaling the weighed spectra. In these circumstances, the elimination of the samples made during the PCA exploratory analysis (E13, E104, M12 and M17) was not found to be appropriate anymore, probably because the number of samples used to compute separate individual PCA models (A, E and N) became too small. Better results were obtained by including in the training sets of these individual models all the spectra of the A and E compounds from the initial training set. The optimization process improved significantly the overall quality of the classification system, and yielded a total correct classification rate of 98.75% , as determined using the number of compounds correctly classified with a significance level of 5% using Coomans' plot.

All tested amphetamine analogs listed in Table 1 were correctly classified with a significance level of 5% (true positives, classification code A*). The system classified 1-phenylethylamines such as α-phenylethylamine (APEA) and N-methyl-α-phenylethylamine (MAPEA), as A amphetamine analogs. These compounds have recently been found in the circuit of illicit drugs in Belgium. Both the pharmacological activity and toxicity of these compounds are virtually unknown, as they were only very recently reported in the context of drug abuse. Thus, they were not considered to be false A positives, as they were already identified in several powders seized during various law enforcement operations, as well as in biological samples from a couple, who were known as drug users, found dead in their apartment[14]. Only one false positive was found, prolintane (1-(α - propylphenethyl)pyrrolidine), which has the N atom included in a heterocycle.

All the tested sympathomimetic amines were correctly classified (see Table 2) with a significance level of 5% (true positives, classification code E*) and no false positive was found. With the exception of prolintane, all negatives were correctly classified (true negatives) as well, and no false negative was found. Cocaine was correctly classified as a negative with a significance level of 5% (classification category N*), as no PCA model was computed for this sample.

Table 3. Hallucinogenic amphetamines recognized among the negatives (classification category A) with their spectral library entry code, acronym, and substituents.

	ID	Name of the compound	Acronym	R$_1$	R$_2$	R$_3$
1.	D20	4-bromo-2,5-dimethoxy-phenylethylamine	2C-B	H	H	H
2.	D37	2,5-dimethoxy-phenylethylamine		H	H	H
3.	D64	4-iodo-2,5-dimethoxyamphetamine	DOI	H	H	-CH$_3$
4.	D79	3,4-methylenedioxyamphetamine	MDA	H	H	-CH$_3$
5.	D82	3,4-methylenedioxy-N-ethylamphetamine	MDEA	-CH$_2$-CH$_3$	H	-CH$_3$
6.	D85	1-(3,4-methylenedioxyphenyl)-2-butanamine	BDB	H	H	-CH$_2$-CH$_3$
7.	D87	3,4-methylenedioxymethamphetamine	MDMA	-CH$_3$	H	-CH$_3$
8.	D106	N-methyl-1-(3,4-methylenedioxyphenyl)-2-butanamine	MBDB	-CH$_3$	H	-CH$_2$-CH$_3$
9.	D113	N,N-dimethyl-3,4- methylenedioxyamphetamine	MDMMA	-CH$_3$	-CH$_3$	-CH$_3$
10.	D139	3,4,5-trimethoxyamphetamine		H	H	-CH$_3$
11.	D145	2,5-dimethoxyamphetamine	DMA	H	H	-CH$_3$
12.	D152	4-bromo-2,5-dimethoxyamphetamine	DOB	H	H	-CH$_3$
13.	D160	4-methylthioamphetamine	4-MTA	H	H	-CH$_3$

Another important advantage of the system is that it discriminates even among the negatives. For example, all substituted phenylethylamines (3,4-methylenedioxyamphetamines, 2,5-dimethoxyamphetamines) with hallucinogenic activity that were tested were correctly discriminated from A amphetamines (nonsubstituted phenylethylamines). However, their structural similarity with A amphetamines was recognized (classification category A), as their identity points in the multidimensional space are closest to the A model (had the smallest sample-to-model distance). The same result was obtained for one of the newest designer drugs, 4-methylthioamphetamine (4-MTA), as shown in Table 3. These hallucinogens were the only compounds in the validation set classified in this manner. Thus, in the case of an unknown, the assignment of the classification category A is a good indication with a high probability that the unknown *is* an amphetamine analog, only with a different substitution pattern than A stimulants. These results are very important from a structure-activity relationship, as the hallucinogens mentioned above also possess the stimulant activity of the A amphetamines.

Another positive characteristic of the system was the very low number of substances doubly classified with a significance level of 5%. Although the closest “hit”, with the smallest sample-to-model distance and leverage, has yielded a correct answer in 98.75% of the trials, 1.25% of the (160) classified compounds indicated a double classification with a 5% significance level.

4. Conclusions

A knowledge-based system was built by performing an exploratory data analysis based on PCA, followed by SIMCA classification. The combination of these methods allowed the development of a tailor-made knowledge-based classification system, based on the vapor phase FTIR spectra of the main ergogenic aids abused by athletes. One of the main data preprocessing steps was the calculation of a discriminant feature weight, which made the samples form well defined clusters during the PCA exploratory analysis. The latter proved the feasibility of discriminating the amphetamine analogs and the ephedrine analogs from their counterparts and among themselves. The information provided by the feature weight combined with the analysis of the score and loading plots allowed the identification of the spectral regions with the most important discrimination power. The SIMCA classification further assigned the class identity to a validation set containing the spectra of a large variety of toxicologically relevant drugs and precursors found in powders or tablets. The classification yielded in its optimized form a total correct classification rate 98.75% with a prediction accuracy adequate for investigations in analytical toxicology and doping control (5% significance level).

Amphetamines and sympathomimetic amines have very specific FTIR spectra. The analytical challenge in identifying novel analogs of these stimulants with FTIR is not related to selectivity, but mainly to the recognition of class similarity, as these stimulants have low weight molecules and small structural differences often yield spectral modifications. The PCA exploratory analysis, as well as the SIMCA validation process, has proved that the system is highly specific, as it assigns the A* class identity compounds *only* to those molecules characterized by the basic skeleton of amphetamine analogs, i.e. a phenyl ring linked by an aliphatic side chain with one (1-phenylethylamines) or two (2-phenylethylamines) carbon atoms, to an amino group. At the same time, E* class identity is assigned only to sympathomimetic amines (ephedrine analogs having the same molecular skeleton as the modeled amphetamines, only with a hydroxy group linked to the first carbon atom of the side chain instead of an alkyl group). In conclusion, the system acting as a screening test for amphetamines found in tablets and powders is more specific than the traditional ones, for which the interference of other substances with the analysis of drugs of abuse is a frequent concern. For example, it is well-known that homogenous enzyme immunoassays yield false positive methamphetamine results for specimens containing high concentrations of ephedrine. The relative specificity of enzyme immunoassays to racemic mixtures, or their cross-reactivity with 3,4-methylenedioxyamphetamine (MDA) or 3,4-methylenedioxymethamphetamine (MDMA) are also matters of concern. Recent developments in cryogenic sample deposition interfaces have improved the sensitivity of the GC-FTIR technique to a degree that it is now even applicable to the low concentrations encountered in the forensic testing of biological fluids[5−7]. It is reasonable to believe that this computer-aided system could be extended to model a traditional screening test in doping control, and obtain results similar to those described in this paper with biological samples by using spectra obtained with cryogenic sample deposition interfaces.

The high selectivity is illustrated also by the fact that the system discriminates not only the modeled classes, but also among amphetamines according to their substitution pattern. The latter is valuable information for correlations between the molecular structure of amphetamines and their pharmacological activity (stimulant or hallucinogenic drugs of abuse) or toxicity. It is worth pointing out that although positional isomers among amphetamines may be discriminated based on mass spectra (MS), the information about the substitution pattern of the phenyl ring is lost during the fragmentation process, and thus a query may only address the entire class of amphetamines[15].

Another important advantage of the system is its modularity. In the case of this system, the specificity of the recognition of the hallucinogenic amphetamine analogs indicates that the introduction of additional rules for the screening of other classes of compounds of toxicological interest (controlled substances such as hallucinogens and narcotics) might be successful. The introduction of additional models for the classes of compounds that were already observed as a cluster in the multidimensional space, outside the boundaries of the present models, might even yield a better characterization of this space and increase the number of negatives classified with a significance level of 5% (classification category N*). However, the introduction of additional models into the classification system is not always beneficial, and thus the effect of additional models remains to be explored.

The automated system that we presented is a fast and user-friendly analysis, involving a small number of steps. Spectra may be exported in Lotus-WK1 format from the FTIR system to a personal computer by using the Perkin Elmer *IRDM*® software. The hardware requirement is a 400-MHz computer, equipped with an Intel Pentium® II processor. The feature weighing may be performed using *Excel*® (Microsoft, USA). The basic chemometric software package *The Unscrambler*® (Camo AS, Norway) is commercially available.

Acknowledgments

This study was carried out as part of the projects BIL 97/75 and BIL 00/39 included in and financed by the Bilateral Scientific and Technological Cooperation Agreement between Flanders and Romania.

References

1. Cordis focus 139, 4, 2000.
2. S.B. Karch, “Drug Abuse Handbook”, CRC Press, New York, 1998.
3. C. Gibson, (Ed.) and C.J. Lyles, “Drugs of Abuse”, Diane Publishing, Upland, 1997.
4. A.K. Cho, (Ed.) and D.S. Segal, “Amphetamine and Its Analogs: Psychopharmacology, Toxicology, and Abuse”, Academic Press, New York, 1994.
5. K.S. Kalasinsky, B. Levine, M.L. Smithand, and G.E. Platoff, Crit. **Rev. Anal. Chem. 23,** 441-457, 1993.
6. K.S. Kalasinsky, B. Levine, and M.L. Smith, **J. Anal. Toxicol. 16,** 332-336, 1992.
7. K.S. Kalasinsky, B. Levine, M.L. Smith, J. Magluilo, and T. Schaefer, **J. Anal. Toxicol. 17,** 359-364,1993.
8. I. Dirinck, E. Meyer, J. Van Bocxlaer, W. Lambert, and A. De Leenheer, **J. Chromatogr. A 819,** 155-159,1998.
9. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Long, P.J. Lewi, and J. Smeyers-Verbeke, “Data Handling in Science and Technology: Handbook of Chemometrics and Qualimetrics”, Part A, Amsterdam, Elsevier, 1997.
10. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Long, P.J. Lewi, and J. Smeyers-Verbeke, “Data Handling in Science and Technology: Handbook of Chemometrics and Qualimetrics”, Part B, Amsterdam, Elsevier, 1998.
11. M. Praisler, I. Dirinck, J. Van Bocxlaer, A. De Leenheer, and D.L. Massart, **Anal. Chim. Acta 404,** 303-317, 2000.
12. E.J. Hasenoehrl, J.H. Perkins, and P.R. Griffiths, **Anal. Chem.64,** 656-663,1992.
13. E.J. Hasenoehrl, J.H. Perkins, and P.R. Griffiths, **Anal. Chem. 64,** 705-710, 1992.
14. J. F. Van Bocxlaer, W.E. Lambert, L. Thienpont, and A.P. De Leenheer, **J. Anal. Toxicol. 21,** 5-11, 1997.
15. M. Praisler, I. Dirinck, J. Van Bocxlaer, A. De Leenheer, and D.L. Massart, (Pattern recognition techniques screening for drugs of abuse with gas chromatography-Fourier transform infrared spectroscopy) accepted for publication in Talanta-The International Journal of Pure and Applied Chemistry.