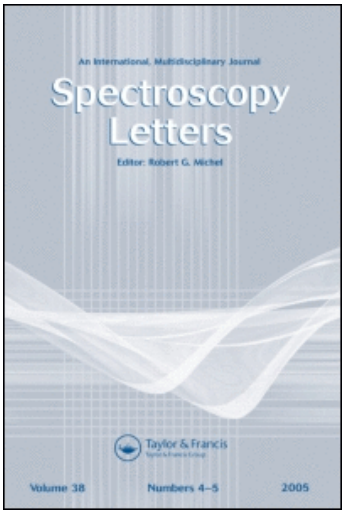


Spectroscopy Letters
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597299

New Strategy of Mass Spectrum Simulation Based on Reduced and Concentrated Knowledge Databases

BoTao Fan[a]; HaiFeng Chen[ab]; Michel Petitjean[a]; Annick Panaye[a]; Jean-Pierre Doucet[a]; HaiRong Xia[b]; ShenGang Yuan[b]
[a] ITODYS, CNRS UMR 7086, Université Paris 7, Paris, France [b] Key Laboratory of Computer Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, People's Republic of China

To cite this Article Fan, BoTao , Chen, HaiFeng , Petitjean, Michel , Panaye, Annick , Doucet, Jean-Pierre , Xia, HaiRong and Yuan, ShenGang(2005) 'New Strategy of Mass Spectrum Simulation Based on Reduced and Concentrated Knowledge Databases', Spectroscopy Letters, 38: 2, 145 — 170
To link to this Article: DOI: 10.1081/SL-200049577
URL: http://dx.doi.org/10.1081/SL-200049577

Spectroscopy Letters, 38: 145–170, 2005

ISSN 0038-7010 print/1532-2289 online
DOI: 10.1081/SL-200049577

New Strategy of Mass Spectrum Simulation Based on Reduced and Concentrated Knowledge Databases

BoTao Fan, HaiFeng Chen, Michel Petitjean, Annick Panaye, and Jean-Pierre Doucet
ITODYS, CNRS UMR 7086, Université Paris 7, Paris, France

HaiFeng Chen, HaiRong Xia, and ShenGang Yuan
Key Laboratory of Computer Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, People's Republic of China

Abstract: A new strategy for mass spectrum simulation is proposed based on the analysis of different reported techniques and developed software packages. This new strategy consists of using four pivot knowledge databases: functional groups, fragmentation pathways, end-point and pseudo end-point fragments, and peak-intensity relationships. The key database is the cleavage mode database, which was constructed by data mining of large mass spectra databases. An important advantage of this concentrated database is its rich information but largely reduced size compared with the databases used in other reported systems. Based on this new strategy, a mass spectrum simulation system, MASSIS, was developed, and the performance was evaluated by comparing the simulated spectra with experimental data for a large population of molecules. The results show that this system possesses high performance. The average ratio of simulated spectra with respect to experimental ones is superior to

Received July 5, 2003, Accepted June 29, 2004

This paper was invited as a contribution to a special issue of the journal concerning Chemical Spectroscopy in China. It was presented at Sichuan University, Chengdu, China, August 1–7, 2004: the 4th Conference for Worldwide Young Chinese Chemists.

Address correspondence to BoTao Fan, ITODYS, CNRS UMR 7086, Université Paris 7, 1, rue Guy de la Brosse, 75005, Paris, France. E-mail: fan@paris7.jussieu.fr

90% for all compounds taken together. The reported mass spectral simulation system in this paper could be the first general software for organic chemistry use.

Keywords: Concentrated knowledge databases, Mass spectrometry, Strategy of simulation, Software package

INTRODUCTION

Computer-aided generation of the mass spectrum for a target molecular structure is always a complicated problem to be solved. To our knowledge, there is presently no system which can give complete solutions. DENDRAL was the first structure elucidation system based on mass spectral data.[1] One of the important routines integrated in this system is devoted to the simulation of mass spectra for candidate structures. This achivement marked the first but extremely important step in mass spectrum prediction. DENDRAL is therefore cited always as a basic reference in this area.

A number of systems have been developed using mass spectral information, such as ACD/Specmanager,[2] AMDIS,[3] BioToolBox,[4] MS Biotools,[5] CombiLab,[6] GPMAW,[7] IDExpert,[8] NIST MS Search Program,[9] MASCOT,[10] MASSFRONTIER,[11] MassLib,[12] SpectralDB,[13] and so forth. All these systems are oriented to structure elucidation. The mass spectra are used only to identify the structural primitives. In some systems, a mass spectrometry database is available, which is used for a double purpose: extraction of structural information and comparing the spectra of candidate structures with those of similar structures stored in the database. In the second case, the best structural match is sought for validating the best candidate.

An important contribution to purely mass spectrum simulation was reported by Gasteiger et al.[14–18] The system MASSIMO[17] developed by Gasteiger et al. is capable of computing the mass spectra for given molecular structures. In this system, some reaction rules are applied to target molecules for calculating the fragmentation scheme and reaction yield and then converting them into simulated mass spectrum. The performance of this system depends largely on the knowledge acquisition process by other software and a reaction database. MASSIMO often gives very good results, but its use seems to be limited by its dependence on other routines, and by the exploitation of original data stored in large reaction databases.

ANALYSIS OF DIFFERENT STRATEGIES

Although the approaches for mass spectrum simulation are poorly reported, we can analyze the different strategies used in reported works.

Searching in Spectral Databases

In fact, this kind of method cannot be considered strictly as a real simulation method for mass spectrometry. It consists only of collecting experimental spectral data, organizing them in databases, and exploiting the spectral information through data mining techniques. BioToolBox is such a software package for characterizing proteins and peptides from mass spectral data.[4] Another convivial system is GPMAW, which is used to interpret peptide and protein mass spectra data[7] by exploiting the famous SWISS-Prot and EMBL databases. These two huge databases gather more than 800,000 protein structures and associated information, including mass spectra.

Both BioToolBox and GPMAW software packages provide a relatively important protein database. It is possible to search these databases to localize the information or mass spectra. In general, this method is highly reliable, because the search is done according to the criteria of a structural match between the target molecule and that stored in the database. Inaccuracies can exist, because the spectra collected in the database might have been obtained under different experimental conditions. Moreover, the results may not be unique, because for the same molecule, several spectra obtained from different experimental resources can be collected in the same database. The exploitation of these databases provides a powerful tool to users, but in fact, such software can only be considered to be data mining tools, not mass spectral simulation systems.

In the same category, another piece of software, MassLib, is nearer to a spectral simulation system. It tries to provide the mass spectrum for a target molecule by means of a database, but in the case where no exact match is available, advanced searching algorithms are applied to extract meaningful mass spectral information based on molecular similarity. However, like the BioToolBox and GPMAW systems, inaccuracies can exist for the same causes.

Artificial Intelligence Methods

Two typical systems in this category are DENDRAL[1] and IDExpert.[8] DENDRAL is an interactive software package that uses the formula, spectrographic data, and encoded heuristic knowledge of organic chemists and geneticists. It applies a data-driven approach to the molecular structure search and explores possible molecular organization in the search for a structure hit. IDExpert is in fact an expert system for structure elucidation. It can give an estimation of the molecular weight by exploring low-resolution mass spectral data. Artificial intelligence techniques have successfully been applied to structure elucidation based on mass spectral information, but their uses in mass spectral simulation are poorly reported. In DENDRAL, the development of mass spectral simulation routines for candidate structures

was limited in its knowledge. This approach was only used to generate the molecular structure based on mass spectra.

An important software package in this category is MassFrontier.[11] This is a complex software package for the management and interpretation of mass spectral data. A unique expert system was implemented in this software for the prediction of fragmentation pathways. This system can support EI, PCI, ESI, and APCI techniques (in MS and MS(n) configuration). We can observe that this system is not only a simple searching procedure in databases, it is more intelligent and can simulate the fragmentation pathways via an expert system.

Exploitation of Fragmentation Reactions

Generally speaking, the mass spectrum results from a series of organic reactions under the conditions of electron impact. If we can reconstruct the fragmentation pathway by a set of reactions, all fragments produced during these reactions constitute the peaks of the mass spectrum. The software MASSIMO is a system exploiting the fragmentation pathways. It applies some reaction rules to a target molecule for calculating the fragmentation scheme and yields of fragments and then converts them into a mass spectrum.

In fact, it is difficult to show a clear separation between artificial intelligence methods and the exploitation of fragmentation reactions; for example, in the DENDRAL system the authors also proposed exploiting the fragmentation reactions.[1]

In the MASSIMO system, for calculation of reaction probabilities, a large number of diverse parameters is needed for evaluating the yields. Lack of certain parameters, even only one, could lead to an erroneous result.

In order to overcome the difficulties cited above and to develop a convivial mass spectrum simulation system with high performance, we proposed a new strategy based on reduced and concentrated knowledge databases.

PRINCIPLE

The general mass spectrometry procedure is shown in Fig. 1. For a target molecule, it is possible to have several fragmentation modes dependent upon the molecular structure, in particular, the functional groups contained in this molecule. Each fragmentation pathway can generate a set of fragments (one or several fragments). These fragments could undergo further fragmentation. This procedure will be stopped when all the final fragments cannot undergo further fragmentation. These final fragments are called end-point fragments. This simple analysis indicates that the simulation of a mass spectrum for a given molecule involves at least three important types of identifiers: (1) recognition of functional substructures included in the target molecule; (2)

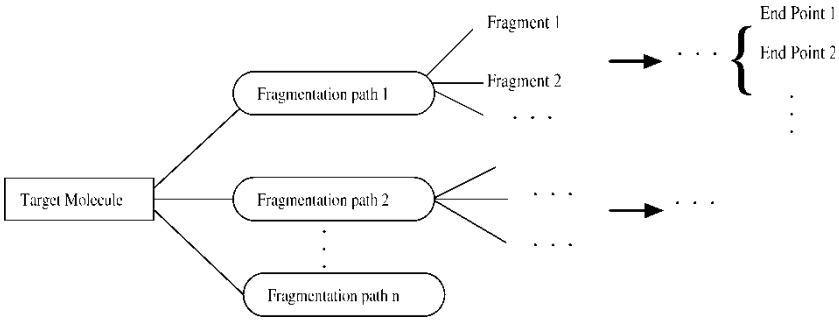


Figure 1. General mass spectrometry procedure.

determination of possible fragmentation pathways associated with functional groups; and (3) identification of end-point fragments. The new strategy of simulation system is based on these identifiers, and the software, called MASSIS (mAss spectrum sImulation system), was developed with this philosophy. Instead of storing large mass spectral databases, we used reduced and concentrated knowledge databases extracted from the literature and obtained from data mining of spectral databases. Three principal databases and an additive pivot databases were constructed. The additive database is used to determine the intensities of simulated mass spectral peaks. We describe these database and their uses in the following sections.

Functional Group Database

Because of the close relationships between functional groups and fragmentation modes, it is better to preclassify the target molecules and then process the computation with a good orientation in other pivot databases. To do this, a pivot database of functional groups has been built according to the criterion of priority. For a functional group starting by a carbon atom, we consider the priority, in order, the nature of the connected atoms, the hybridization of the atom, and the nature of the bond. This example is shown in Fig. 2.

All molecules are classified into 26 principal categories, and they are further classified as 277 subcategories.

Fragmentation Mode Database

This is the most important pivot database. It was constructed by data mining of a huge mass spectral database (provided by the laboratory of mass spectrometry, Shanghai Institute of Organic Chemistry, with about 120,000

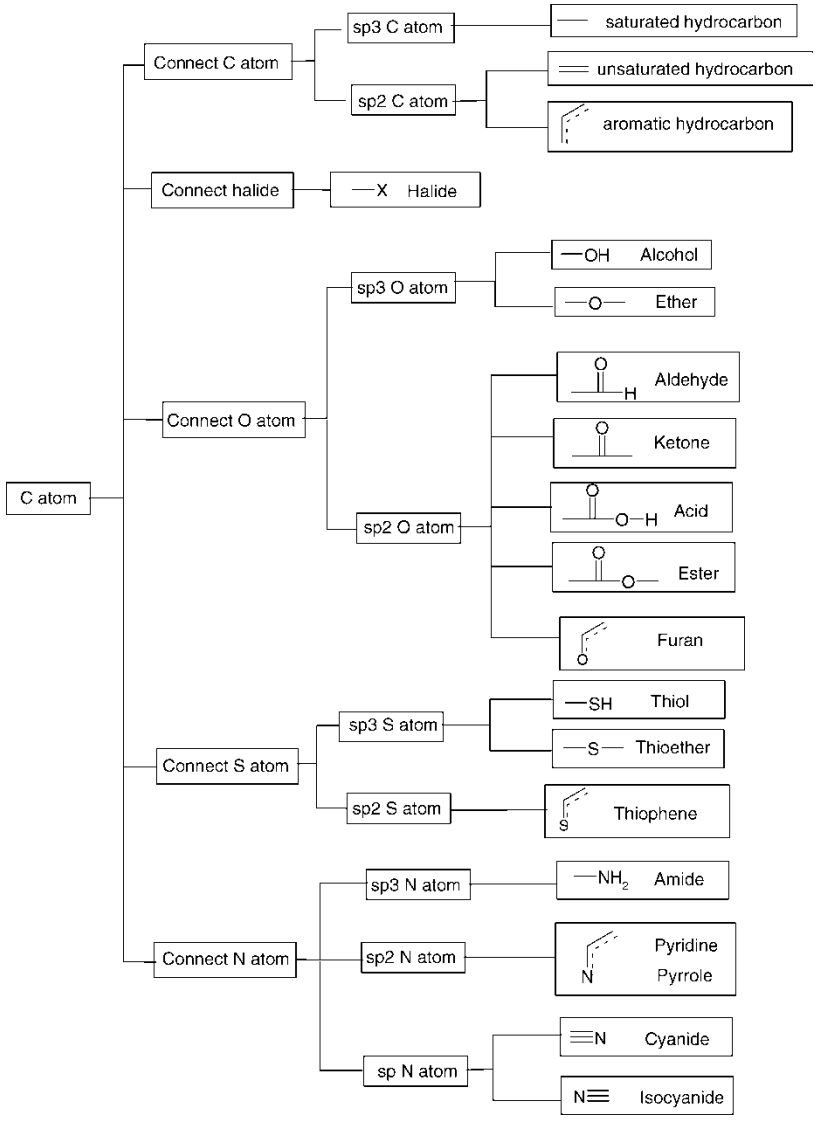


Figure 2. Functional group classification starting from carbon atom.

spectra available) and extraction from the literature.[19] In some software packages, like GPMAW or BioToolBox discussed above, the searching is carried out in whole original spectra databases. In MASSIMO, developed by Gasteiger et al., the reaction databases are used. Differently from these approaches, we tried to extract the fragmentation mode information from

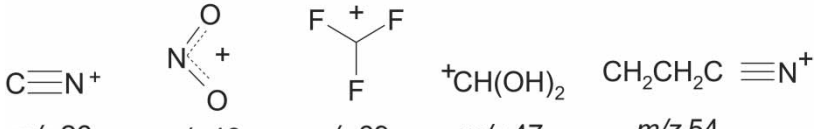
the mass spectral data and to find the corresponding cleavage reactions from the literature. The cluster analysis has been used to classify all available spectra according to the functional groups. One molecule containing several groups can be arranged in several classes. Then for each class of molecules, careful analysis of their corresponding spectra leads us to deduce the possible fragmentation modes, while the most possible modes were compared with those extracted from the literature in order to determine the cleavage rules. Another statistical analysis was done simultaneously for deriving the relationships of peak intensities and environmental influence. This will be further discussed in the section "Peak Intensity Database."

Actually, we gathered 1625 cleavage rules covering 26 types of organic compounds, corresponding to 26 categories in pivot database of functional groups. The categories and corresponding numbers of rules, as well as examples for each category, are shown in Table 1. We should underline that this simulation system is an open one. New rules can be added easily in a user-defined pivot database. This permits us to later extend the system to a larger range of applications.

Compared with the use of original and rough data of mass spectra and computationz on reaction databases, our fragmentation mode database is in fact a reduced and concentrated knowledge database. The information on mass spectra and fragmentation pathways is compressed in a size-reduced database. This pivot database is the kernel of this new strategy for mass spectral simulation. It covers a good part of the usual organic molecules, except metal-containing compounds.

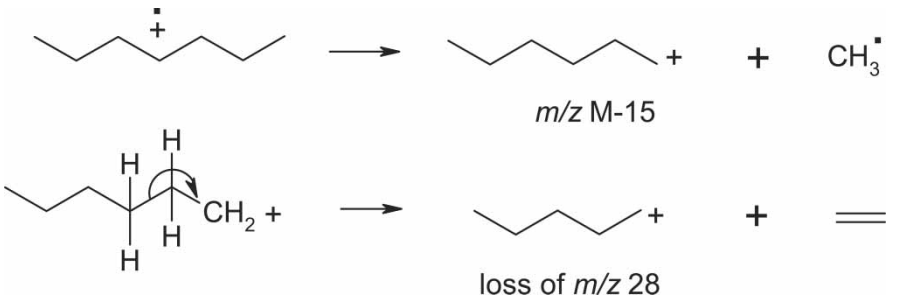
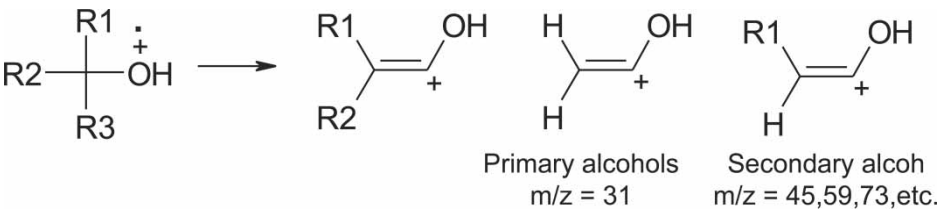
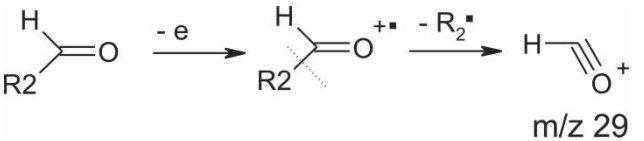
Database of End-point Fragments and Pseudo End-point Fragments

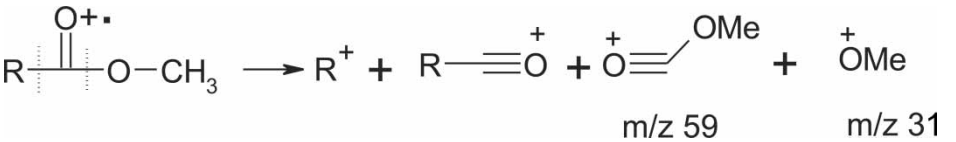
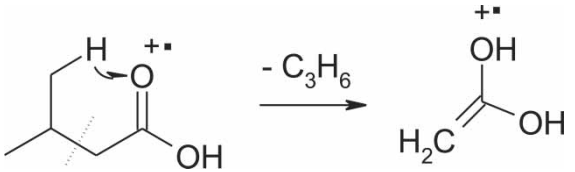
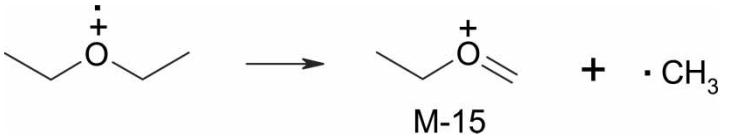
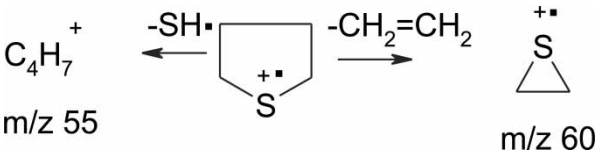
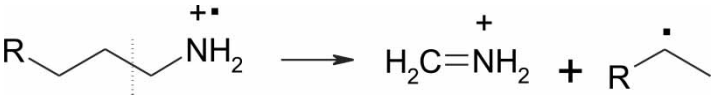
In fact, this database contain some small fragments that can be classified into two categories: end-point fragments and pseudo end-point fragments. An end-point fragment means that the fragment cannot undergo further fragmentation. When such a fragment is detected, the system stops the treatment for this fragment. In fact, 89 end-point fragments have been collected in the pivot database; some examples are shown below:



In contrast, a pseudo end-point fragment can undergo further fragmentation, but the pathway of the further fragmentation is well defined. The advantage to separation of this kind of fragment from ordinary cleavage database is to

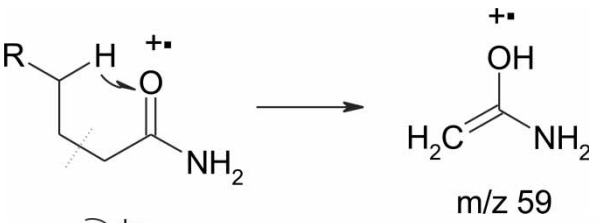
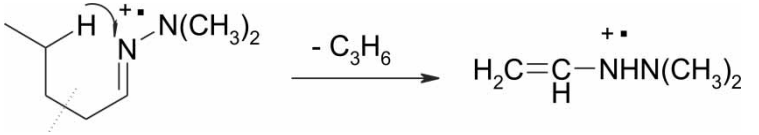
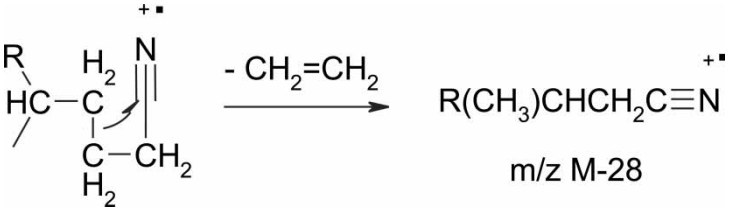
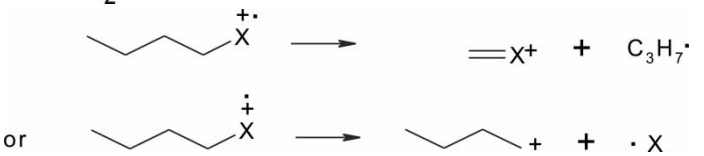
Table 1. Categories of cleavage rules for functional compounds

No.	Functional compounds	Number of rules	Example
1	Hydrocarbon	68	
2	Alcohol	49	
3	Aldehydes and ketones	96	

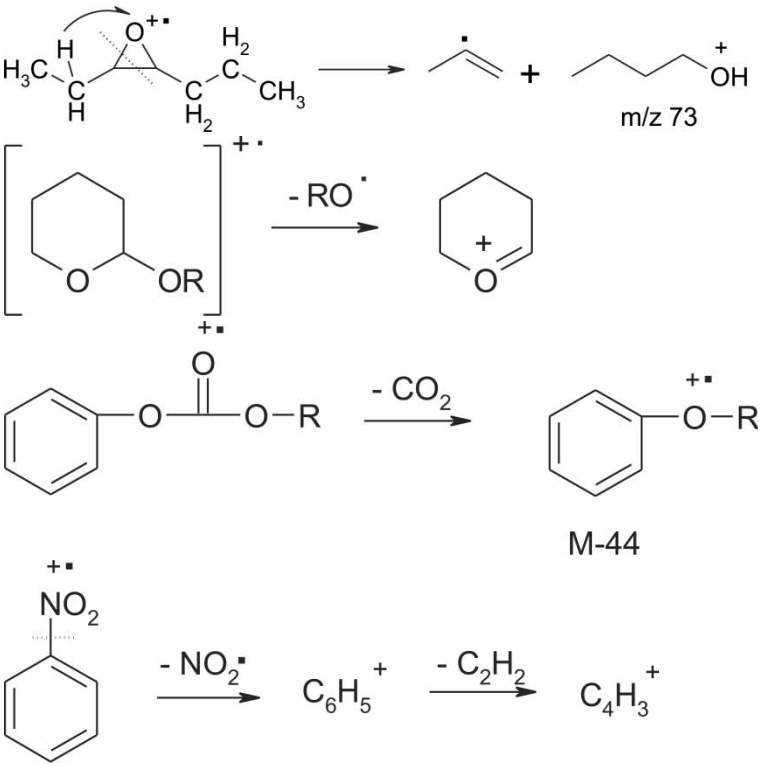
4	Esters and lactones	120	
5	Acids and anhydrides	52	
6	Ethers, acetals, ketals, and orthoesters	128	
7	Thiols and thioethers	56	
8	Amines, N-oxides, and nitrosamines	106	

(*continued*)

Table 1. Continued

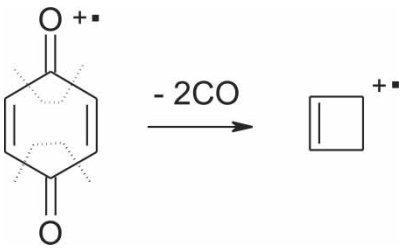
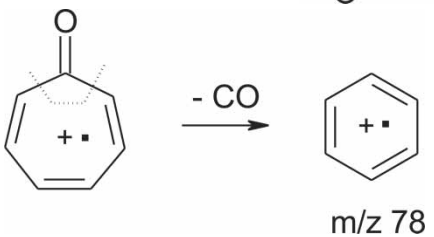
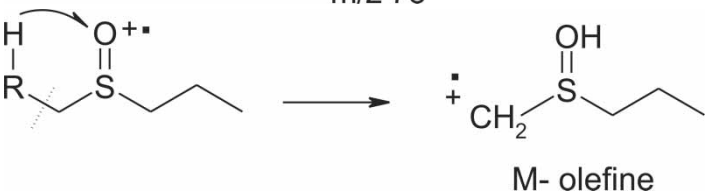
No.	Functional compounds	Number of rules	Example
9	Amides and lactams	104	
10	Oxime, hydrazone, and nitrophenylhydrazone	136	
11	Cyanides	112	
12	Halides	25	

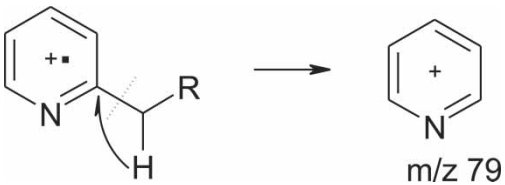
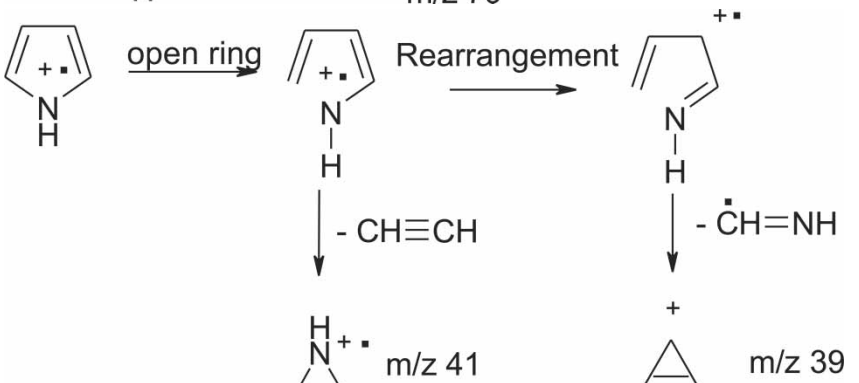
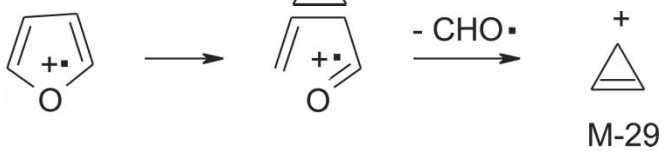
13	Epoxides	35
14	Alcohol derivatives	20
15	Carbonates, carbamates, and ureas	98
16	Nitro compounds	56



(*continued*)

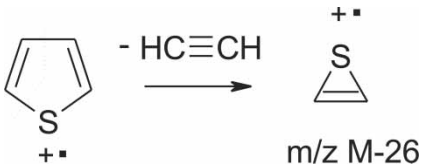
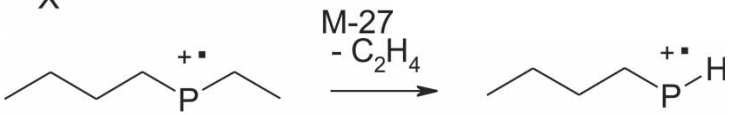
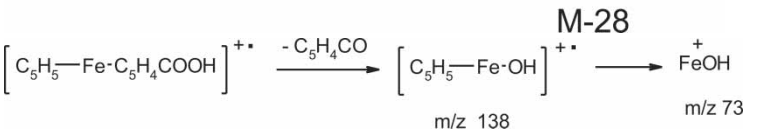
Table 1. Continued

No.	Functional compounds	Number of rules	Example
17	Quinones	28	
18	Tropones and tropolones	31	
19	Sulfoxides and sulfones	70	

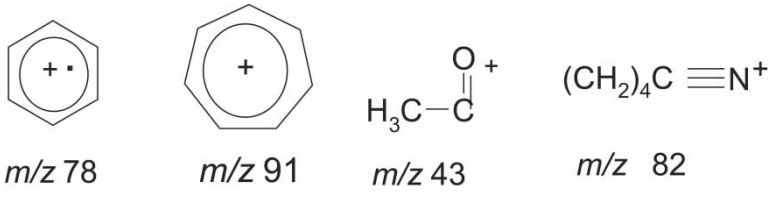
20	Pyridines, quinolines, and isoquinolines	62	
21	Pyrroles and indoles	25	
22	Furan and related compounds	34	

(*continued*)

Table 1. Continued

No.	Functional compounds	Number of rules	Example
23	Thiophenes and benzothiophenes	33	
24	Thiazoles, oxazoles, oxadiazole, imidazoles, and pyrazoles	38	
25	Organic phosphorus	19	
26	Organometallic compounds	23	

save computation time. When such a fragment is examined, the system gives directly the fragments that issue from the fragmentation pathway without checking the fragments after each step. For example, the following fragments can be classified into this category:



These two types of small fragments are identified by means of an additional index that orients the system to apply directly to the defined cleavage mode if the pseudo end-point fragment is detected, otherwise, the system stops the procedure for the end-point fragment.

Peak Intensity Database

This is an additive pivot database for assigning the intensities to simulated peaks. Simulating the intensity of a peak with good accurency is always very difficult. Using energetic criteria to determine the favorability of bond dissociation on electron impact is too complex. Gasteiger et al. used a set of parameters to calculate the yields and probabilities of fragments. This procedure involves the computation task for each fragment and a number of parameters predefined. The peak intensity is an important parameter in a mass spectrum, but it is not primordial compared with the fragments. By applying data-mining techniques to huge databases, we can determine statistically the relationships between peak intensities and the environmental influence around the considered fragment. This implies the determination of the environment for a fragment. We use the concept of atom-focused FREL (fragment reduced to environment that is limited) to describe this environment.[20,21] An example is given in Fig. 3. In our system, we consider the layers until a topological distance of 4, because the layers superior to 4 have practically no effect on intensity.

To save storage memory, each fragment was encoded as a string. For example, for the fragment shown in Fig. 4, if the bond C1–N is broken, the positive charge is localized on atom C1, which is an sp^3 hybridization carbon (C3) with connectivity 3 (d3) and it is not included in a ring (r0). This atom is now selected as a focus, with encoding string (C3d3r0). The first layer environment of C1 contains an N atom with sp^3 hybridization and connectivity 3, a C atom with sp^3 and connectivity 1, and a C (cyclic atom) with hybridization sp^2 and connectivity 3. The encoding string for the first

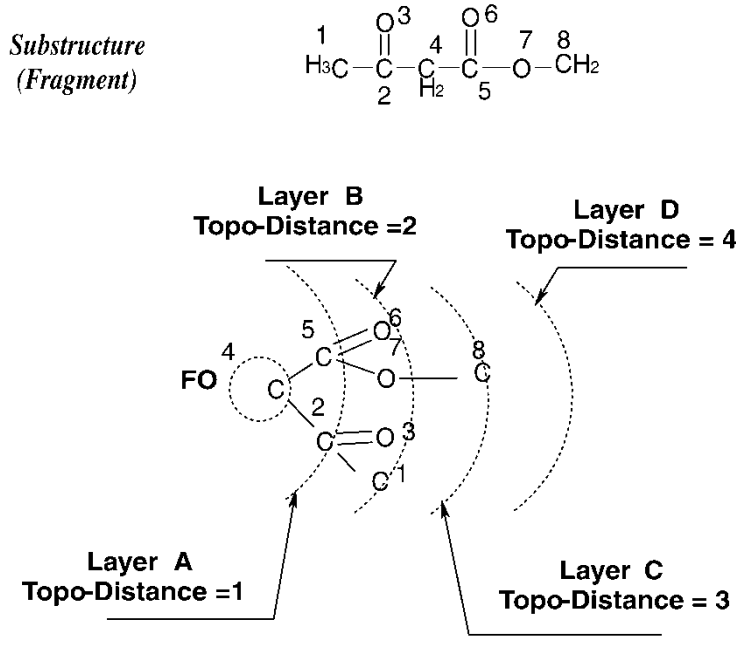
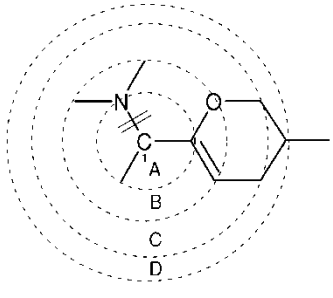


Figure 3. Example of atom-focused FRELs.

layer is therefore (N3d3r0C3d1r0C2d3r1). The same analysis can be applied to other layers. Finally, we obtain the whole encoding string for this fragment by assembling the strings of all layers (including the focus) (see Fig. 4).

We extracted all fragments from the mass spectral database and grouped the same fragments together. Then the average intensity of the corresponding



Focus Layer A Layer B Layer C Layer D

C3d3r0/N3d3r0C3d1r0C2d3r1/O3d2r1C3d1r0C3d1r0C2d2r1/C3d2r1C3d2r1/C3d3r1

Figure 4. Encoding an atom-FREL.

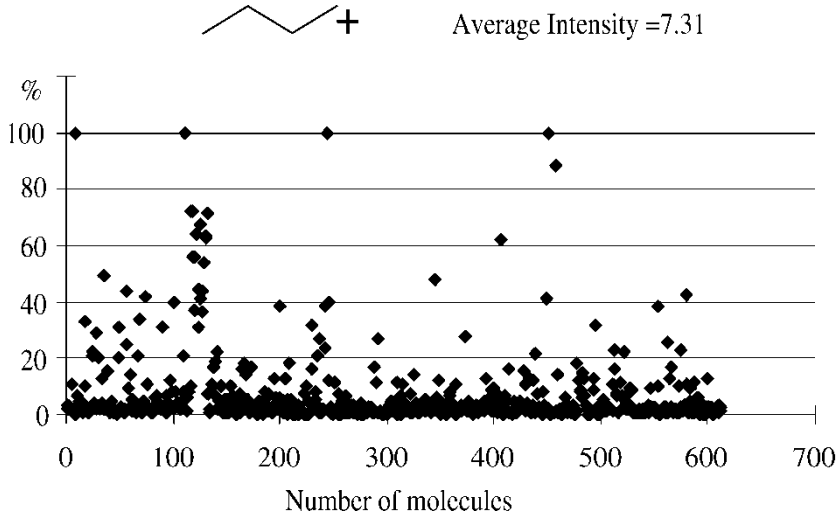


Figure 5. Average intensity calculation for fragment $CH_3(CH_2)_2\ CH_2^+$.

peak was been calculated. An example is shown in Fig. 5. This result is consistent with the previous study from the mechanism point of view.[22]

Finally, this pivot contains about 50,000 fragments with the intensities determined according to their environments.

PROCEDURE

The procedure of the MASSIS system can briefly be illustrated by the flowchart shown in Fig. 6.

The process is executed through four principal steps corresponding to uses of four concentrated pivot databases of knowledge.

Step 1: The input structure is checked to determine its principal functional groups by using Ullmann's backtracking algorithm,[23] a generally used algorithm for searching graph isomorphism. It is often the case where a molecule contains several functional groups. The system marks every functional group by predefined flags, and each function will be treated according to the priorities defined in the pivot database of functional groups.

Step 2: The target molecule is then oriented to an adequate location of the cleavage mode database according to the functional groups, thanks to the linkage index between the pivot databases. As the environment of a functional group is described with the FREL concept around the central atom, this corresponds therefore to a shortest-path problem. Dijkstra's

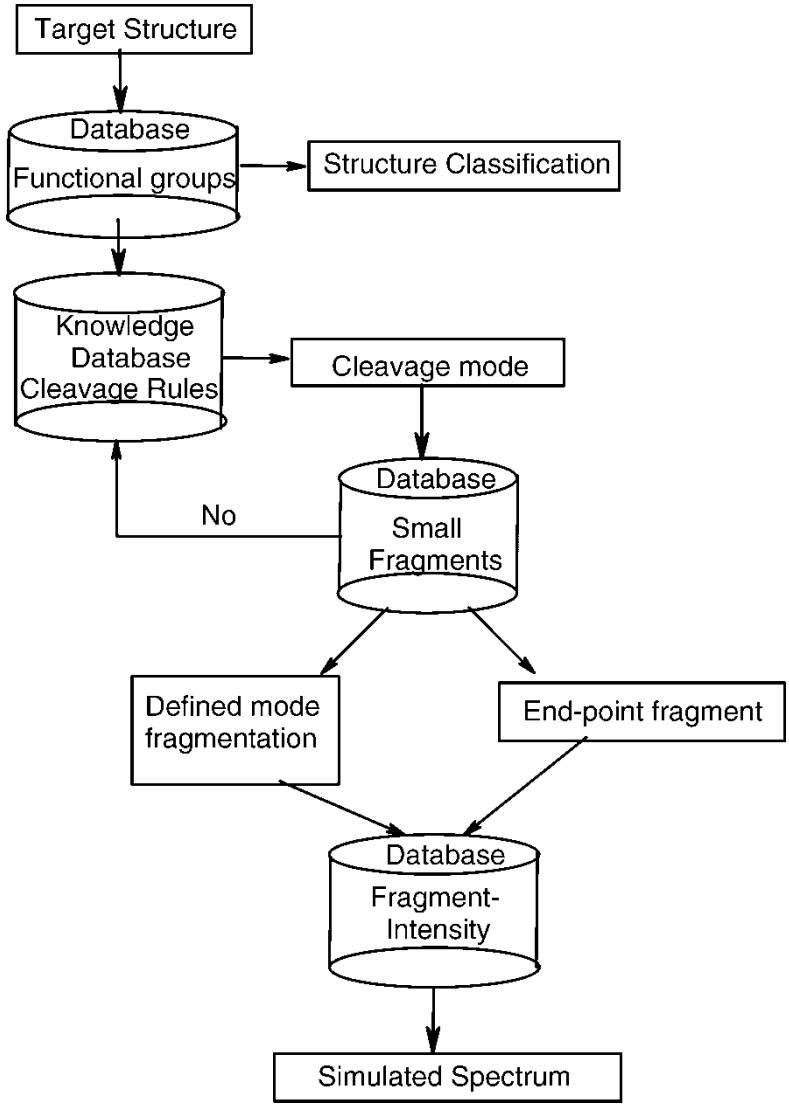


Figure 6. Flowchart of MASSIS simulation system.

algorithm[24] can be applied as an efficient technique to solve this shortest-path problem. In Fig. 7, as an example, we show how the system generates the fragment m/z 60 via McLafferty rearrangement.

Step 3: The generated fragments are then compared with the end-point fragments and pseudo end-point fragments stored in the third pivot

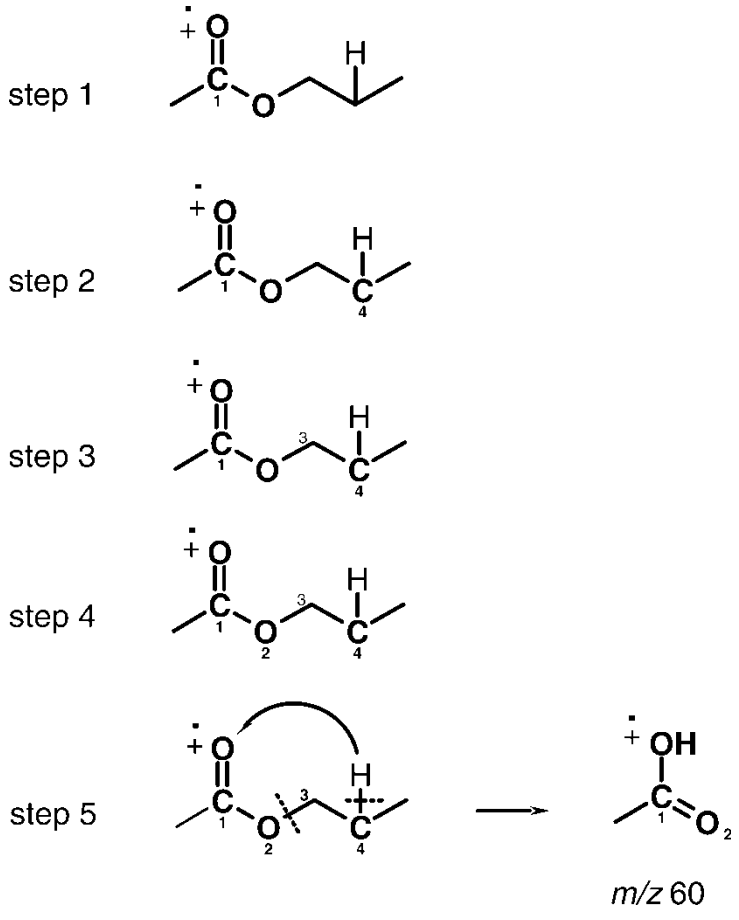


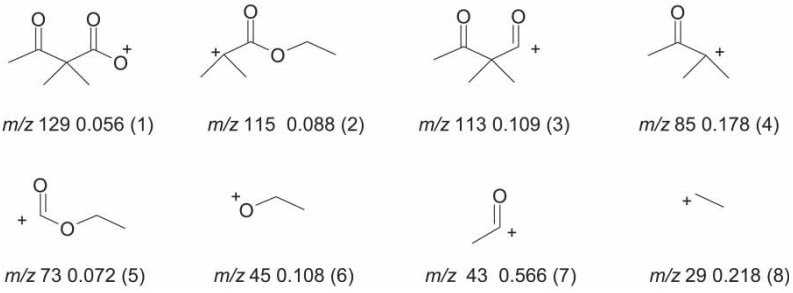
Figure 7. Example of fragment generation.

database. Two substeps should be considered: (1) if the examined fragment is an end-point fragment, it is then recorded and the system stops further treatment; (2) If it is only a pseudo end-point fragment, the system searches in an auxiliary database for the corresponding subfragments that should be produced via a defined cleavage pathway starting from the query fragment and records these new fragments instead of the pseudo end-point fragment.

The steps 1–3 are repeated until the whole target molecule is computed; all generated fragments are now recorded as a set of encoding strings.

Step 4: In this step, the system determines the peak intensities by means of the peak-intensity relationship database. For example, the EI mass spectrum of

2,2-dimethyl-3-oxo-butyric acid ethyl ester contains 8 fragments:



The corresponding encoding strings are:

Peak 1: O3d2r0 C3d2r0 C2d3r0 C3d1r0 O2d1r0 C3d4r0 C3d1r0 C3d1r0 C2d3r0 O2d1r0 C3D1r0

Peak 2: C3d4r0 C2d3r0 C3d1r0 C3d1r0 C2d3r0 O2d1r0 C3d1r0 O3d2r0 O2d1r0 C3d2r0 C3d1r0

Peak 3: C2d3r0 O3d2r0 O2d1r0 C3d4r0 C3d2r0 C3d1r0 C3d1r0 C2d3r0 C3d1r0 O2d1r0 C3d1r0

Peak 4: C3d4r0 C2d3r0 C3d1r0 C3d1r0 C2d3r0 O3d2r0 O2d1r0 O2d1r0 C3d1r0 C3d2r0 C3d1r0

Peak 5: C2d3r0 C3d4r0 O3d2r0 O2d1r0 C3d1r0 C3d1r0 C2d3r0 C3d2r0 O2d1r0 C3d1r0 C3d1r0

Peak 6: O3d2r0 C2d3r0 C3d2r0 O2d1r0 C3d4r0 C3d1r0 C3d1r0 C3d1r0 C2d3r0 O2d1r0 C3d1r0

Peak 7: C2d3r0 C3d4r0 O2d1r0 C3d1r0 C3d1r0 C3d1r0 C2d3r0 O3d2r0 O2d1r0 C3d2r0

Peak 8: C3d2r0 O3d2r0 C3d1r0 C2d3r0 O2d1r0 C3d4r0 C3d1r0 C3d1r0 C2d3r0

By searching these code strings in pivot database, we can determine the intensity for each peak.

PERFORMANCE

The performance of our system was studied by comparing simulated spectra with experimental data. A set of 7370 molecules have been tested. An example is given in Fig. 8. The corresponding fragmentation pathway is shown in Fig. 9, and the numeric values are gathered in Table 2.

If we consider only the peaks with intensity $>5\%$ for this structure, there are 17 peaks with intensity $>5\%$ in the experimental spectrum; all these peaks

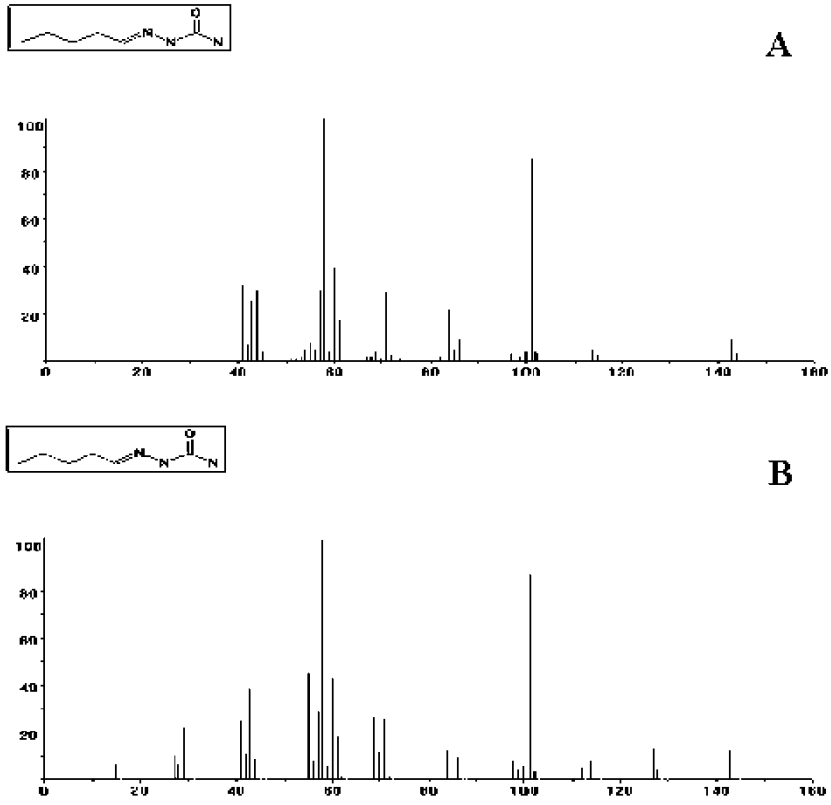


Figure 8. Comparison of mass spectrum of *N*-valeraldehyde semicarbazone. (A) Experimental spectrum; (B) simulated spectrum.

have been found in the simulated spectrum. The ratio of the simulated spectrum to experimental data is 100%. We can also remark that the residuals between the intensity of experimental and simulated spectra are relatively small, except for the fragments m/z 44 and m/z 55 (see Table 2). But these two peaks are not the most important ones (see Fig. 9).

An extensive comparison between experimental and simulated mass spectra has been carried out with 7370 structures. This test set consists of:

89 boron-atom-containing compounds;

1 beryllium compound;

280 metal-containing molecules;

917 hydrocarbons (without hetero atom);

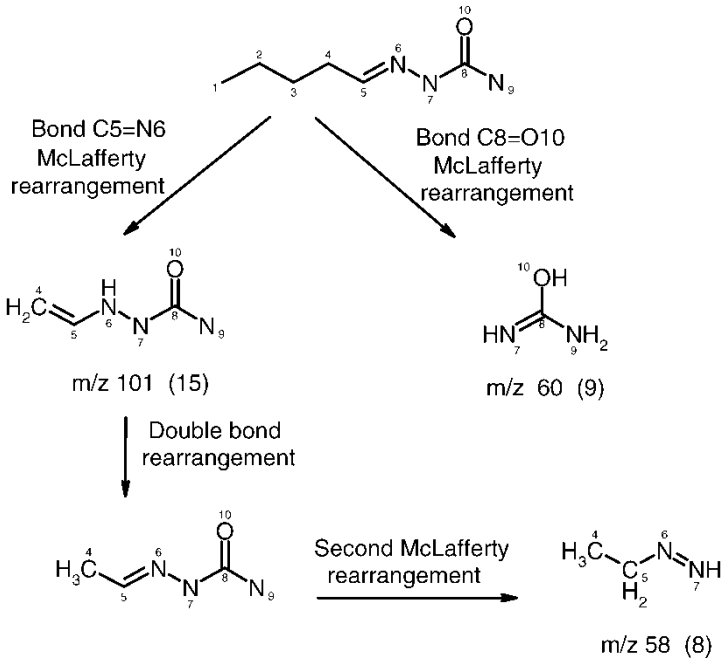


Figure 9. Analysis of bond cleavages.

Table 2. Comparison of experimental and simulated mass spectra[a]

No.	m/z	Experimental intensity (EI)	Simulated intensity (SI)	Fragment formula	Cleavage mode	Residual of EI and SI
1.	15	0.0670	0.1633	CH3 +	Simple cleavage	−0.0963
2.	27	0.4580	0.1228	C2H3 +	Simple cleavage	0.3352
3.	28	0.0950	0.0661	C2H4 +	Chain cleavage	0.0289
4.	29	0.4850	0.4900	C2H5 +	Simple cleavage	−0.0050
5.	39	0.2290	0.0408	C3H3 +	Lose hydrogen	0.1882
6.	41	0.5630	0.2049	C3H5 +	Simple cleavage	0.3581
7.	42	0.2560	0.1083	C3H6 +	Chain cleavage	0.1477
8.	43	0.9990	0.9900	C3H7 +	Simple cleavage	0.0090
9.	55	0.1190	0.0931	C4H7 +	Simple cleavage	0.0259
10.	56	0.2650	0.1375	C4H8 +	Chain cleavage	0.1275
11.	57	0.4740	0.3675	C4H9 +	Simple cleavage	0.1065
12.	70	0.1780	0.1123	C5H10 +	Chain Cleavage	0.0657
13.	71	0.4510	0.2940	C5H11 +	Simple cleavage	0.1570
14.	100	0.1260	0.1200	C7H16 +	M +	0.006

[a]Only the peaks with intensity >5% have been taken into account for comparison.

2630 compounds with monofunctional groups;

3453 compounds with multi-functional groups.

Among the molecules bearing mono- or multifunctional groups, we have the following partition:

Alcohols:	1383	Aldehydes and ketones:	930
Esters:	1006	Acids:	291
Ethers:	2186	Thiols:	822
Amines:	2682	Amides:	469
Cyanides:	170	Other N-atom-containing compounds:	2099
Epoxides:	47	Phosphorus compounds:	54

Other functional groups.

It should be pointed out that if a molecule contains several functional groups, it will be accounted in each category.

The comparison between experimental data and predicted spectra show a good performance of our MASSIS system. The statistical result for 7370 test molecules is shown in Fig. 10 and Table 3. From Fig. 10, we remark that for some tested structures, the simulated spectra are not satisfactory. A careful examination shows that these structures are essentially the metal-containing compounds and boron-atom-containing compounds. In the first case, the

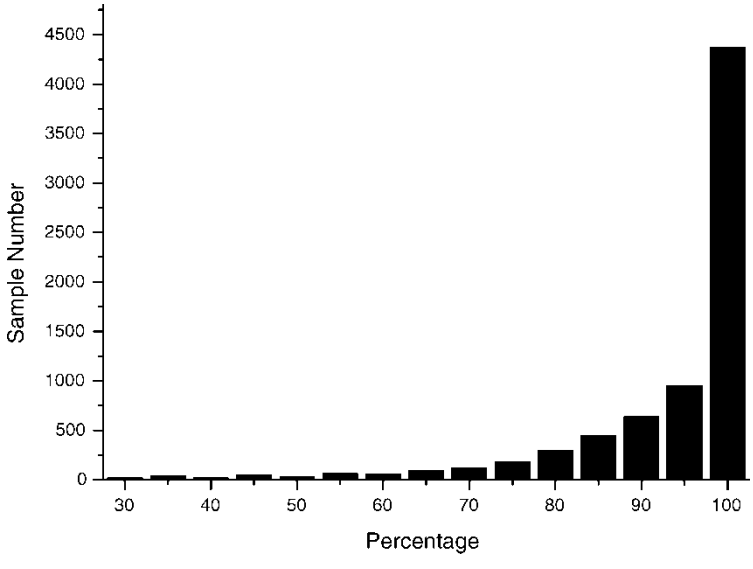


Figure 10. Statistical histogram of test molecules. Percentage: ratio of simulated peaks in respect with experimental data; Sample number: number of molecular structures.

Table 3. Statistical study of simulated mass spectra

Percentage	Sample number	Percentage	Sample number
25–30%	19	65–70%	122
30–35%	37	70–75%	181
35–40%	18	75–80%	295
40–45%	51	80–85%	444
45–50%	27	85–90%	637
50–55%	66	90–95%	952
55–60%	59	95–100%	4370
60–65%	92		

fragmentation rules gathered in the database are largely insufficient for covering all metal-containing compounds, which leads to poor prediction ability of system for organometallic compounds. In another case, the boron-atom-containing molecules have not been taken into account in our fragmentation rule databases. The system cannot simulate their mass spectra correctly.

In Table 3, the percentages were evaluated by matching the simulated peaks with experimental ones without consideration of intensity. The overpredicting peaks were not taken into account. As indicated by a referee, using this criterion for assessing the quality of performance is not a very good measure, but before finding an adequate criterion, it can show at least a global measure based on statistics of a large population.

If we take x_i the mean percentage value for an interval (for instance, $x_i = 0.975$ if the percentage interval is 95–100%), the average ratio of simulated spectra on experimental data can be evaluated as

$$\begin{aligned}\bar{R} &= \frac{\sum_i f_i x_i}{\sum_i f_i} \\ &= \frac{19 \cdot 0.275 + 37 \cdot 0.325 + 18 \cdot 0.375 + \cdots + 952 \cdot 0.925 + 4370 \cdot 0.975}{7370} \\ &= \frac{6691.8}{7370} = 0.9080\end{aligned}$$

It means that the ratio of simulated peaks with respect to experimental data might be superior to 90%. If we neglect the metallic complexes or metal-containing molecules, the ratio of simulated peaks on experimental data for usual organic compounds could be superior to the average value calculated above.

CONCLUSIONS

Starting from the analysis of traditional strategies used in mass spectral simulation, we proposed a new strategy and developed the MASSIS system

for mass spectral simulation based on reduced and concentrated knowledge databases. Four pivot databases have been constructed by applying data-mining techniques to large mass spectra databases. The key database is the cleavage mode (fragmentation pathway) database, which is concentrated in information but with a largely reduced size compared with usual mass spectra databases.

The system MASSIS was developed around these four databases linked by inter-indices. The procedure includes four principal steps: functional group classification, fragmentation according to the presence of functional groups, end-point or pseudo end-point fragment check, and determination of peak intensity. Finally, a graphic display routine permits us to visualize the simulated spectrum. This system was developed in Visual C++ 6.0 language with Visual Interface under Microsoft Windows 98/2000/XP.

The performance was evaluated by comparing the simulated spectra and experimental data. A statistical study has been carried out with a large population. The average ratio of simulated spectra with respect to experimental ones is superior to 90% for all compounds taken together. If we consider only the usual organic compounds, this ratio could be ameliorated. This indicates that MASSIS possesses high performance as a mass spectrum simulation system. We hope this can be the first available mass spectrum simulation software that can be used independently (it need not be associated with mass spectral databases and other software or routines).

ACKNOWLEDGMENTS

The authors are thankful to the Embassy of France in China (grant no. 2001995), the National Centre of Scientific Research of France (CNRS, France, grant no. 12475 and no. 14916), the Chinese Academy of Sciences, the Minister of Science and Technology of China (grant no. 2002AA231011 and no. 2003CB114401), and the National Natural Science Foundation of China (grant no. 29573148 and no. 29872048) for the financial support of this work. We thank also the referees for their helpful suggestions and very interesting discussion.

REFERENCES

1. (a) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. *Applications of Artificial Intelligence for Organic Chemistry. The Dendral Project*; McGraw-Hill: New York, 1980; (b) Buchanan, B. G.; Feigenbaum, E. A. DENDRAL and Meta-DENDRAL: their applications dimension. *Artif. Intell.* **1978**, *11*, 5–24; (c) Lederberg, J. How DENDRAL was conceived and born. In *ACM Symposium on the History of Medical Informatics*; In *A History of*

Medical Informatics; Blum, B. I., and Duncan, K., Eds.; Association for Computing Machinery Press, New York, 1990; pp. 14–44. National Library of Medicine, 1987; (d) Luger, G. F.; Stubblefield, W. A. *Artificial Intelligence and the Design of Expert Systems*; Benjamin/Cummings Publishing, Redwood City, California, 1989.

2. Available at http://www.acdlabs.com/.
3. Available at http://www.amdis.iaea.org/amdis.htm.
4. Available at http://www.appliedbiosystems.com/.
5. Available at http://www.bruker-daltonik.de/.
6. Available at http://www.massspec.com/.
7. Available at http://www.welcome.to/gpmaw.
8. Available at http://www.asitinc.com/Hyperspectral.htm.
9. Available at http://www.nist.gov/.
10. Available at http://www.matrixscience.com/.
11. Available at http://www.chemsw.com/.
12. Available at http://www.masslib.com/.
13. Available at http://www.galactic.com/.
14. Bauerschmidt, S.; Hanebeck, W.; Schulz, K.-P.; Gasteiger, J. Elucidation of reactions in the mass spectrometer. *Anal. Chim. Acta* **1992**, *265*, 169–181.
15. Schulz, K. P.; Bauerschmidt, S.; Höllering, R.; Gasteiger, J. Automatic elucidation of reactions in the mass spectrometer. In *AIP-Conference Proceedings 330: E.C.C.C.1, Computational Chemistry*; Bernardi, F., Rivail, J. L., Eds.; American Institute of Physics: Woodbury, NY, 1995; pp. 726–733.
16. Gasteiger, J.; Hanebeck, W.; Schulz, K.-P.; Bauerschmidt, S.; Höllering, R. Automatic analysis and simulation of mass spectra. In *Computer-Enhanced Analytical Spectroscopy*; Wilkins, C. L., Ed.; Plenum Press: New York, 1993; Vol. 4, pp. 97–133.
17. Hanebeck, W.; Rafeiner, K.; Schulz, K.-P.; Röse, P.; Gasteiger, J. Towards the automatic generation of a mass spectrum from the structure of a compound. In *Software-Development in Chemistry 4*; Gasteiger, J., Ed.; Springer-Verlag: Heidelberg, 1990; pp. 187–195.
18. Gasteiger, J.; Hanebeck, W.; Schulz, K. P. Prediction of mass spectra from structural information. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 264–271.
19. McLafferty, F. W. *Interpretation of Mass Spectra*; University Science, New York; Sausalito, California 1980.
20. Dubois, J. E.; Laurent, D.; Viellard, H. Système DARC. Principes des recherches de corrélations et equation générale de topo-information. *C.R. Acad. Sci. Paris Ser C* **1967**, *264*, 1019–1022.
21. Dubois, J. E.; Mercier, C.; Panaye, A. DARC topological system and computer aided design. *Acta Pharm. Jugosl.* **1986**, *36*, 135–169.
22. Ning, Y. C. *Organic Structural Identification and Organic Spectrometry*; Science Press: Beijing, 2000.
23. Ullmann, J. R. An algorithm for subgraph isomorphism. *J. Assoc. Comput. Machin.* **1976**, *23*, 31–42.
24. Dijkstra, E. A note on two problems in connexion with graphs. *Numerische Mathematik* **1959**, *1*, 269–271.