Computer-Assisted Structure Elucidation, Ranking of Candidate Structures, Based on Comparison between Predicted and Observed Mass Spectra.
Tomas H. Varkony, et al., 1977

Computer Assisted Structure Elucidation, Ranking of Candidate Structures, Based on Comparison Between Predicted and Observed Mass Spectra.

Tomas H. Varkony, Raymond E. Carhart, and Dennis H. Smith,

Stanford University
Department of Chemistry, Computer Science, Genetics
Stanford, California 94305

Mass spectral data may be used in structural studies in several ways: 1) For example, we can create a fragmentation theory based on examination of sets of known structures and their associated mass spectra. 2) Or, assuming that peaks in the mass spectrum of an unknown originate from unrearranged molecular ions, we can propose possible structures by combining the fragments together under the guidance of a fragmentation theory. 3) When the structure is given we can predict a mass spectrum which will obey the rules of a fragmentation theory.

All of these operations can, in principle be translated into a set of instructions in a computer program. Some of these programs we have presented in the past. One of these programs uses a combination of a predictor which uses a theory of mass spectrometry to predict the spectra of candidate structures, and an evaluation function which compares the predictions with the observed spectrum of the unknown, assigning a goodnes-of-fit score to each candidate. The candidates are sorted based upon how well the predicted spectra match the observed.

1) Allbreaks: The most general theory assumes that every bond in the molecule can be broken, and the predicted spectrum is obtained by the calculation of the composition (or the mass) of combinations of all connected pieces of the molecule. This theory is too general and not very useful in structure elucidation problems. We can constrain this theory in different ways: 1) The binary mode, which is in use in the part of the program called MSPRED. In this mode a break is permitted or forbidden by a set of constraints defined by the chemist and there is no difference in the evaluation of two different process which utilize different permitted cleavages. 2) In the "weighted" mode in MSRANK we attach different values for the constraints used to allow or to forbid certain cleavages. 3) When the unkown is related to a previously investigated class of compounds, we can use subgraph rules to describe desired process in predicting the spectrum.

MSPRED: This program uses a theory ("half-order" theory) which assumes that every bond in the molecule can be broken under certain constraints. In predicting the spectrum, in the binary mode, MSPRED explores all possible cleavages of the molecule within these general (user-defined) constraints. Constraints include limitation of the number of bonds broken and the number of steps in a process, the proximity of pairs of bonds (i.e. whether or not two adjacent bonds can break in a given process) the multiplicity or aromaticity of each cleaved bond, the

allowed hydrogen atoms transfered from or into the charged fragment and the neutral fragments which can be lost. The program calculates the composition and the mass of the fragment which can be obtained in a fragmentation process. The program then combines these results into a predicted mass spectrum with peaks of uniform intensity. The best predictive theories of mass spectrometry are limited to families of closely related structures (i.e., class specific theories). However, given the wide variety of structural types which can be produced by CONGEN (Ref 1) and REACT (Ref 2), it is necessary for MSPRED to use this very general model of mass spectral fragmentation.

Evaluation and Ranking: For this general approach we decided to use an evaluation function which takes into account that peaks at high m/e values and high intensity have more diagnostic value then peaks in the low m/e region of the spectrum with low intensity. The simplest form of various evaluation functions we have used is given in Equation 1.

SCORE = M^Si^RI^Si^R 1

where M^Si^R is the mass of a peak present in both the predicted and observed spectra. I^Si^R is the intensity of the (correctly predicted) peak in the observed spectrum. We expect the half order theory to be overly complete in the sense that, when applied to the correct structure for an unknown, it will doubtless predict many plausible fragments which are not observed. This simply reflects the fact that the "break everything" approach to mass spectrometry is a considerable oversimplification.

Example: Our interest in mass spectra of steroids led us to examine a class of mono-keto androstanes as a test case. We obtained the high resolution mass spectra for 10 of the 11 possible mono-ketoandrostanes. These 11 structures were our list of candidate structures. We predicted the high resolution spectra for each of the 11 structures using the half-order theory, and then ranked them against each of the 10 observed spectra. In most of the cases (Table I), the correct structure was ranked first and in the remaining it was ranked second. The half-order theory is insufficient to differentiate among monoketo- androstanes when the keto group is located in one of the 4 possible positions in ring A or among structures which differ in the location of the keto substituent in Ring D. We are now doing a systematic study of various classes of compounds by ranking the spectrum of a known structure against a CONGEN or REACT generated list of structures which contains the correct structure among several which are closely related. In most of the test cases (including low and high resolution mass spectral data) the correct structure was ranked among the upper ten percent of the structures. We are optimistic that the results of ranking based on the half-order theory can be used as a preliminary filter to divide a set of candidate structures into two portions, one of which has an extremely high probability of containing the correct structure. To this set of top-ranked structures we can apply a more detailed fragmentation theory to make specific predictions.

Table I. RANKING OF MONOKETO-ANDROSTANES (Half order theory)

STRUCTURE	RANKING	STRUCTURES BETTER RANKED (keto position)	SCORE	WITH THE SAME STRUCTURES
				7 2
		6 16 1		17,15 11 2
		12 3 1		1,2,4 17 2
	15,16	2,3,4,1 6	1 12	1 15 1
17,16 1 1		2,3,4 4 1	1,2,3	

A modification of the theory used for MSPRED is based on the following principle: of two candidate structures for an unknown, the most likely structure is the one which explain the observation most "simply" - i.e., with the fewest complex explanations involving many bond cleavages and the transfer of many hydrogen atoms. The evaluation function used by this program, which we call MSRANK, is based on a quantitation of this principle.

In predicting a spectrum, MSRANK like MSPRED explores all possible cleavages of the molecule with constraints similar to the user-defined constraints used in MSPRED. These constraints, as in MSPRED, include the number of bonds broken and the number of steps in a process, the proximity of pairs of cleaved bonds (i.e., whether or not two adjacent bonds can break in a given process) and the multiplicity or aromaticity of each cleaved bond. Within these general limits, the user also supplies numerical plausibilities from 0 to 1 on the various kinds of breaks which are allowed to occur. For example, he might give unit plausibility to 1-bond cleavages, .8 to 2-bond processes and .6 to 3-bond processes. Aromatic-bond, multiple-bond and adjacent-bond cleavages, if allowed, are given separate plausibilities, as are the allowed neutral transfers. MSRANK combines these values multiplicatively in evaluating the overall plausibility of a specific mass spectral process, and that value is associated with the corresponding predicted mass point. If two different processes predict the same mass point, the highest plausibility value is retained. The result is a predicted spectrum with numbers attached to each peak, interpreted roughly as the "reasonable-ness" or "simplicity of prediction" measure. A "reward" is given to every observed peak which is predicted, the amount being proportional to the plausibility of the prediction and (at the user's option) to the intensity and/or mass value of the observed peak. The sum of rewards for all observed peaks then constitutes the overall score for the candidate which gave rise to the predicted spectrum.

We used MSPRED and MSRANK for ranking isomers of Methyl-esters of several aromatic acids (some of them obtained from urine extracts), by comparing their predicted spectrums to the observed low resolution spectrum of the methyl-ester. For example; we generated the 19 isomers of Hippuric-acid- methyl-ester attaching to the different positions of the aromatic benzene ring the combination of a methyl-ester group, a nitrogen atom, a carbonyl and an additional carbon atom. The constraints used by MSPRED were insufficient to rank the correct structure in the highest position although it was ranked among the first four structures. Using the same type of constraints but with attached plausibility values, improved the ranking, and MSRANK ranked the correct structure in the highest position.

2) Rule-based theory. When the candidate structure is known to belong to a previously investigated class of compounds, then we can use additional information to predict a more precise mass spectrum. This information is in the form of specific fragmentation rules. These rules are described by a subgraph, a break and related hydrogen or neutral transfers. For every match of the rule's subgraph to a candidate structure, the program calculates the composition peaks, and collects these peaks, for all the applications of the rules, into a predicted spectrum. When predicting mass spectra using a rule based theory, we have found that we can predict a more accurate spectrum and get a better ranking then with the half order theory.

We have investigated the mass spectral fragmentation of a group of macrolide antibiotics. We obtained the high resolution mass spectra of the aglycone part of five 12-membered macrolide antibiotics. We used these data to create fragmentation rules. These rules are the combinations of different single bond cleavages which can cleave the molecule into two pieces. The rules are alpha cleavages to the oxygen substituents, McLafferty rearrangements and beta cleavages to double bonds or oxygens.

All the structures posses the same macrolactone skeleton and they are different in the position of the oxygen substituents and the degree of unsaturation which is present as double bond. We generated the isomers of each of these structures by changing the position of the three oxygens around the macrolactone ring, and allowed them to exist as alcohols or ketones or aldehydes.

When ranking these isomers against the observed high resolution spectra, MSRANK ranked the correct structure at the top half of the candidate structures, but in most of the cases the program could not differentiate among a large number of isomers which get the same score as the correct structure. By using the previously obtained fragmentation rules, we improved the ranking.

This part of the program is still in a development stage. In the future we plan to use information about the frequency of correct application of the rule in the set of compounds from which the rules were developed. We call this information the confidence factor associated with the rule. Other important information we plan to use is the intensity range associated with the peaks which are predicted by a rule.

Summary

We have illustrated a number of approaches to extend the concept of computer-assisted structure elucidation beyond that of simple structure generation. We have illustrated how mass-spectral information together with a computer program can assist chemists in both planning prior to structure generation and, subsequently, testing of candidates. In work described here, the chemist plays an integral part in effective use of the problem-solving tools we provide in the form of interactive programs.