Computer-Based Structure Determination: Then and Now

Morton E. Munk

Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona 85287

Received April 22, 1998

A widely practiced approach to structure elucidation is based on the chemical and/or the spectral properties of the compound of unknown structure. As the power and sophistication of spectrometers improved, spectral data assumed a more central role. With increasing amounts of such data to process, it was natural to look to the computer to enhance productivity. This paper traces the development of computer-based tools for structure elucidation.

INTRODUCTION

The problem of determining the chemical structure of compounds classed as organic is as old as organic chemistry itself. Some organic compounds, which were believed to be associated only with living organisms until the isolation of urea from ammonium cyanate by Fredrich Wöhler in 1828, have been known since earliest antiquity. Although many early distinguished chemists grappled with the problem of determining their chemical structures, it was not until 1859 when August Kekulé, in an intuitive rationalization of data existing at the time, first described the "structure" of simple carbon compounds in terms of a graphical representation of the connectivity and bonding of the atoms in the molecule. With modification, that graphical representation has survived to this day.

Building on the idea of the valence of atoms commonly found in organic compounds, Kekulé was able to rationalize in structural terms the existence of different compounds of the same empirical formula and apply this reasoning to develop the earliest approach to structure determination. Starting with the empirical formula of the unknown, all theoretically possible structural isomers were elaborated. The assignment of the correct structure was then reduced to a problem of distinguishing among the different isomers, a step that, at that time, required a study of the chemical behavior of the unknown and/or a synthesis of one or more of the isomers. Of course, this process was limited to the simplest of compounds and limited by the paucity of chemical information available at the time, but the resemblance of the basic concept to modern approaches to structure determination is noteworthy.

From the beginning, the procedures used in structure determination were empirical in nature and based on observation and experience. For a long time, observed relationships between chemical behavior and structural features formed the foundation of the process. By the middle 1950s, a new and powerful probe of chemical structure was beginning to emerge: molecular spectroscopy. Developments in this area came rapidly, and conventional structure determination became more heavily dependent on spectral data than chemical data as the utility of the diagnostic correlations between spectra and structure became apparent. The more recent availability of large, computer-readable spectral databases has significantly enhanced the value of such correlations in structure determination. Another comparative advantage of spectral data is the speed of their acquisition relative to chemical data.

In the hands of experts, arriving at a correct assignment in conventional structure determination has not usually been a problem. However, as the demand for structure determination increased, it became evident that the time required by the chemist/spectroscopist to fully interpret the collective spectral data and reduce the inferred structural information to a plausible molecular structure was the limiting factor in productivity. Thus, in conventional structure determination, as succeeding generations of spectrometers continued to deliver spectral data faster, productivity, not the correct assignment of structure, became the problem. Given the need to process large amounts of spectral data quickly, it was quite natural for concerned investigators to look to the computer, and indeed, for more than 30 years, the computer has been the centerpiece of numerous efforts to augment the productivity of chemists and spectroscopists engaged in structure determination. The task has not been easy because the goal is an ambitious one: to develop a machine capable of simulating a high level of human intelligence. In pursuing this goal, the techniques of artificial intelligence (AI), a term used here in its broadest sense, provided a fruitful framework for exploration because structure determination, although a large and complex domain, is relatively focused and well-defined. It is also a domain in which there exists a large and cohesive body of knowledge. For these reasons, this problem domain was one of the earliest extensive studies of the application of AI.[1] Early in the work, it became evident that expert system approaches, which attempt to mimic step-by-step what is done by the chemist, were not applicable to all stages of the structure determination process because, as conventionally practiced, that process includes intuitional leaps which are currently not amenable to computer modeling. Thus, different strategies are required for computer implementation which efficiently lead to the same end result.

Structure determination problems are usually of one of two types: structure verification or structure elucidation. In structure verification, the more commonly encountered of the two types, there is enough information available, perhaps based on the use of well-established reactions in a synthetic

10.1021/ci980083r CCC: $15.00
Published on Web 08/01/1998

scheme, for the chemist to propose a probable structure for an intermediate or the final product. However, since unexpected outcomes of reactions are possible, verification of the proposed structure may be required. In the second type of structure determination problem, structure elucidation, the information available to the chemist is insufficient to permit a structure to be proposed; i.e., the structure of the compound is unknown.

There are three major capabilities of central importance in spectral-based structure determination. Spectrum interpretation is the process by which spectral data are reduced to structural inferences which are usually expressed in terms of substructures predicted to be present or absent in the compound under study. Structure generation serves to exhaustively generate all molecular structures compatible with these structural inferences. Spectrum prediction and comparison are important in evaluating their relative probability of being correct.

In structure verification the important capability is spectrum prediction and comparison. (High-quality assigned spectral databases—databases containing not only the structures and spectral data but, in the case of NMR, signal assignments as well—also play a central role.) In contrast, comprehensive computer-based systems for structure elucidation require all three capabilities.[2] The discussion that follows is framed in terms of structure elucidation.

Computer programs that focus only on one of the three described capabilities can serve as useful stand-alone tools for chemists and spectroscopists or they can be incorporated into more comprehensive structure elucidation systems. With few exceptions (e.g., CHEMICS and STREC; see COMPREHENSIVE STRUCTURE ELUCIDATION SYSTEMS), early developmental efforts concentrated on single-capability programs.

STRUCTURE GENERATION

Introduction. Although no two experts practice the art of structure elucidation in exactly the same way, there are some common elements that can be discerned. In the early stages of the problem, the molecular formula is determined and structural information is derived from chemical and/or spectral evidence. The collective structural information is usually expressed as a set of substructures predicted to be present or absent in the unknown. That information, together with its molecular formula, is the usual input in structure generation. If the entered information is considered to be the initial partial structure of the unknown, then structure generation can be described as the expansion of that partial structure into all molecular structures compatible with it. Depending on the information content of the partial structure, expansion may lead to a single structure, very many structures, or somewhere in between. If the number is manageable, the structures themselves can serve as a useful guide for the chemist in designing the most efficient experimental strategy to narrow the choices to one.

Structure generating procedures of practical value should meet three requirements. First, the procedure should be exhaustive. There are numerous examples of compounds with incorrect structural assignments in the primary literature, not because the structure is inconsistent with the evidence but because another equally compatible structure was overlooked.[3] Computers, unlike chemists at times, have no preconceived ideas about plausible structure types. Also, unlike chemists, computers, if properly programmed, excel in accuracy in the performance of repetitive, tedious tasks. Second, the program should execute efficiently, i.e., on a time scale that enhances, not detracts from, user productivity. Third, the output should be a set of irredundant structures.

Although many structure generators have been described, the underlying procedures fall into one of two classes: structure assembly or structure reduction. Experience has shown that each of these approaches has application in structure elucidation problems. Although a compound of unknown structure is not fully characterized until both its constitution and absolute configuration are known, the structure generators which have been described produce only constitutional isomers. Stereoisomer generation and assignment of configuration can follow at a later stage (see **Stereochemistry**).

Structure Assembly. Joining substructural fragments together at their residual bonding sites in all possible ways is the earliest form of structure generation. In the hands of the chemist, this step in the structure elucidation process procedure is the most tedious and least appealing. It is also the step most prone to error. Thus, it is not surprising that some of the earliest work in the area of computer-based structure elucidation has been devoted to the development of structure generators.

The development of structure generators began in the 1960s. Four independent groups saw the need for such a tool at about the same time. For the most part, the resulting structure generators can be described as procedures to systematically search for all valid interconnections between the residual bonding sites on the inferred substructures and on the unaccounted for atoms. The process can be viewed as the expansion of a partial structure to all complete molecular structures compatible with it. The number and diversity of the structures generated is an indication of the information content of the partial structure.

Structure generators that make interconnections between residual bonding sites can be classified as structure assemblers. In the early 1960s two independent studies led to the development of stand-alone structures generators: ASSEMBLE[4] and CONGEN.[5] These two programs possess both similarities and differences. ASSEMBLE is an algorithm for partial structure expansion using connectivity matrices. The process resembles a depth-first tree search. Redundancies are reduced in number by using the topological symmetry of the expanding partial structure to reject redundant connections. As each molecular structure is generated, it is canonically named, providing for the retrospective elimination of any duplicates.[6]

CONGEN follows a two-step procedure. Initially, the structural building units are expressed in a reduced form based on predefined vertex graphs (wherein all “nodes” (atoms) of degree less than 3 are suppressed) and assembled to all possible intermediate structures. Each of these intermediate (reduced) structures is then expanded to all possible compatible complete molecules in a process described as embedding. In a later version of CONGEN, the vertex graph-based first step was replaced with a connectivity matrix-generating algorithm.[7]

The input to both programs is very similar and consists of three parts: the molecular formula, fragments, and constraints. The fragments and the constraints can be chemist- or computer-generated. Fragments, i.e., substructures, serve as the structural building units to be interconnected. In both ASSEMBLE and CONGEN, fragments must have no atoms in common; that is, they cannot overlap one another. This non-overlap restriction presents a problem because the usual procedures for inferring the presence of substructures can produce substructures with identical segments. In many cases, these identical segments may or may not represent overlapping atoms and it is not a simple task to distinguish between the two possibilities. In contrast, constraints, some of which may also be expressed as substructures, are not directly involved in the bond-making process. Instead, the information contained in constraints constrains the bond-making procedure, thereby reducing the solution space that must be explored and the number of plausible structures generated. The non-overlap restriction does not apply to constraints. There may or may not be overlapping segments of atoms. No distinction is required. The output of these structure generators is an exhaustive and irredundant set of molecular structures, each of which is compatible with the input.

In using ASSEMBLE and CONGEN, the substructures predicted to be present can be examined to identify real and potential overlap among them, the latter of which may not be trivial. Only those that with certainty do not possess overlapping segments of atoms are entered as fragments. The remaining substructures predicted to be present and any predicted to be absent are entered as constraints. Thus, the assignment of required substructures as fragment or constraint is a user decision. If it is determined that two or more substructures are potentially overlapping, the one which is richest in structural information is usually selected to be the fragment; others are entered as constraints. The chemist can choose to avoid the step of assessing overlap and merely select the most information-rich required substructure as a fragment and enter all others as constraints. In general, this is not the preferred approach since information entered as fragments is used far more efficiently than information entered as constraints. For more rapid program execution, it is desirable to enter as much of the available structural information as possible as non-overlapping fragments.

ASSEMBLE partially addresses the overlap restriction with a mechanism that can accommodate one-atom overlaps in substructures used as fragments.[8] However, it is the user who must identify the potential one-atom overlaps and enter the information by means of special constraints called atom tags (see below). An extension of CONGEN called GENOA provides a more comprehensive solution to the problem.[9]

Molecule building in GENOA is a two-stage process. Although each stage is a bond-making procedure, their functions differ. In the first stage, the set of potentially overlapping required substructures is processed to express as much of the information as possible as non-overlapping substructures. To initialize the process, the user first selects one of the substructures as a core. A second substructure is then selected by the user, and the program then uses a procedure called a constructive substructure search to generate all possible ways of expressing those two substructures as non-overlapping fragments. Next, a third substructure is selected by the user, and the program again determines all possible ways of expressing those three substructures as non-overlapping fragments. At each step of this process, the user can examine each of the new partial structures and, using human intelligence, choose to retain or delete. The process continues until all substructures have been utilized. The final result is a collection of sets of non-overlapping fragments. If some pruning of resulting partial structures is not done along the way, the number of sets of non-overlapping fragments in the final collection can be quite large.

In the second step of the overall process, GENOA expands each of these sets of fragments into complete molecules, a step which is comparable to structure generation in CONGEN. But, since each set consists of non-overlapping substructures, all substructures in each set can be used as fragments in the final molecule building step. The result is a very efficient process.

The development of workable structure generators advanced more rapidly than computer-based spectrum interpretation tools. Thus, early users of ASSEMBLE and CONGEN/GENOA or their spectroscopists did the interpretation and prepared the input in the required format. In order to maximize the range of structural information that could be given to the structure generators, each included a set of user-entered constraints to limit the number of structures generated accordingly. Those available in ASSEMBLE are illustrative.[8] ASSEMBLE's constraints are of two types. Local constraints provide additional information about the immediate environment of a specific atom in a substructure that is serving as a structural building unit (fragment). Global constraints describe structural features that characterize the unknown compound as a whole.

Local constraints are implemented by adding atom tags to specific atoms within structural fragments. The neighboring atom tag describes an atom immediately contiguous to the tagged atom. The element type is required. Specifying the hybridization of the neighboring atom, the bond by which it is joined to the tagged atom, the number of attached hydrogens, and the minimum and maximum number of such neighboring atoms to be allowed is optional. Information contained within an atom tag is not counted as part of the structural fragment. Thus, the neighboring atom tag is the mechanism referred to earlier that may be used to specify a potentially overlapping atom in a fragment that serves as a structural building unit.

A second tag, the hybridization tag, can specify the hybridization of the tagged atom. The cycle tag sets restrictions on the presence or absence of the tagged atom in a cycle of specified size. The unsaturation tag can require the presence or absence of an unsaturated linkage conjugated to the tagged atom. The branch tag can be used to indicate the presence of a group of atoms of known composition, but unknown structure, attached to the tagged atom. With the vicinal hydrogen tag, the total number of hydrogen atoms allowed on atoms adjacent to the tagged atom can be specified. The internal unsaturation tag forbids bond formation using bonding sites within a fragment, thereby restricting bridging or additional unsaturation.

In contrast to atom tags, global constraints are not site-specific. They provide information about the unknown compound as a whole. There are three unsaturation constraints with which allowed unsaturation equivalents can be specified in terms of an exact number or a range of rings and/or double and triple bonds. Two constraints control cycles. One specifies the number of permitted cycles, either an exact number or range; the other controls the number of cycles of specified size. The number of cycles and the size of the cycles may either be exact or a range. Often, for a specific element—carbon in particular—there may be information on the number of atoms of a particular hybridization and hydrogen multiplicity. One constraint allows the input of such information. The number and types of hydrogens, e.g., cyclopropyl hydrogens, aromatic hydrogens, olefinic hydrogens and hydrogens on carbon atoms bearing electronegative atoms such as oxygen, nitrogen, and halogen, can also be specified. A frequently used constraint requires the presence of a substructure that is not entered as a fragment due to a potential overlap with another substructure used as a fragment. The same constraint can also be used to forbid the presence of a substructure. If the substructure is required, either an exact number or a range may be specified. If more than one occurrence of a substructure is specified, the user can either allow or disallow overlaps between these units in valid structures.

ASSEMBLE also includes a program that predicts the number of signals expected in the ^{13}C NMR spectrum of every molecular structure generated. A generated structure is retained for output only if the predicted number of signals is within a user-set tolerance of the observed number. During structure generation, ASSEMBLE detects the formation of several commonly encountered strained structural features which result in molecular instability. A particular path of structure generation is terminated as soon as such features are constructed. Thus, highly strained structures never appear in the set of valid compounds unless the strain detector is disabled by the user.

ASSEMBLE and CONGEN/GENOA were designed to assemble molecules using any substructure derived by the chemist or a computer program. In contrast, the structure generators that are part of CHEMICS[10,11] and STREC,[12,13] two programs that link spectrum interpretation and structure generation, function to assemble molecules from a predefined set of multiatom fragments. In CHEMICS, these multiatom fragments do not overlap one another and, for a particular set of elements, can lead to the generation of any plausible compound. (In the earliest version of CHEMICS only carbon, hydrogen, and oxygen were considered.)

In its operation, CHEMICS initially predicts the presence or absence of each of the fragments in the library using an automated spectrum interpreter. In general, not all of the surviving fragments are valid. The program next identifies all sets of the surviving fragments that are consistent with the molecular formula of the unknown and any user-entered constraints. Each such set, corresponding to one or more complete molecules, is treated as a separate structure assembly problem. The structure-generating algorithm creates a separate connectivity stack (equivalent to a connectivity matrix) which defines the possible interconnections between the fragments in each set. The procedure then identifies all unique and irredundant ways of joining the fragments together which are compatible with all of the available evidence.

STREC, like CHEMICS, first identifies a surviving set of predefined fragments produced by an automated spectrum interpretation procedure. Subsets of these fragments consistent with the molecular formula and other information are represented in connectivity matrices. A complex two-stage algorithm, resembling that used in CONGEN, then generates all possible permutations of each matrix. Some redundancies and disjoint structures have to be removed retrospectively. The development of two other systems, SEAC[14] and EX-SPEC,[15] which link spectrum interpretation and comparable structure assembly programs, followed in the 1980s.

Structure Reduction. In a sense, structure generation by structure reduction[16] is the reverse of structure assembly. In contrast to structure assembly, in the initial state of a structure reduction problem all possible bonds between a set of atoms are made. This initial problem state is a “hyperstructure” usually corresponding to an extremely large family of isomeric molecular structures, each of which is represented by a subset of the bonds of the hyperstructure. The hyperstructure is expressed as a binary bond adjacency matrix.

The structure generator, COCOA, is an exhaustive, recursive bond-removal procedure which systematically searches the solution space for all valid subsets of bonds in the initial matrix which correspond to molecular structures compatible with the input. Bonds are removed in a systematic way—usually more than one at a time—until a one-to-one mapping between bonds remains, i.e., a complete molecule. The process is repeated until all compatible structures have been produced. In program SESAMI,[17] the information used to guide bond removal is provided by the dual output of a computer-based spectrum interpreter (INTERPRET) which is tightly linked to the structure generator. Using an approach similar to that in CHEMICS, the first track of INTERPRET gives rise to a set of explicitly defined, uniformly sized, atom-centered fragments that are predicted to be present in the unknown. Generally, there are many more invalid than valid fragments. Using the same spectral data as input, the second track of INTERPRET produces one or more predicted substructures which may be of any size and any degree of ambiguity. The output can include alternative interpretations of the spectral data.

Structure Assembly vs Structure Reduction. Conceptually, structure reduction offers a number of advantages over structure assembly. All of the structural information is used prospectively, whereas in structure assembly some of the information is necessarily used retrospectively. As a result, more of the invalid structures are eliminated before they are generated. Structure reduction has no requirement for non-overlapping fragments. It also allows the utilization of alternative interpretations of data prospectively and without any preprocessing. Symmetry information, derived in part from ^{13}C NMR data, can also be used prospectively during structure reduction.

However, in spite of the apparent conceptual advantage of structure reduction, in practice, structure generators based on either of the two methods described have played a significant role in computer-based structure elucidation and will likely continue to so. In comprehensive structure elucidation systems, the choice can depend on the nature and relative importance of the spectrum interpretation and spectrum prediction components.[2]

Stereochemistry. Many compounds possess one or more stereocenters and are isolated as pure stereoisomers. For the chemist, the assignment of the structure of such an compound is not complete until its absolute configuration is identified.

Although programs such as CHEMICS, SESAMI, and STREC, which generate constitutional isomers, can be of great value, they were not originally designed to treat stereochemistry. Extensions of all three programs have now been described; however, the pioneering work in stereoisomer generation was done earlier.[18,19] In this approach, the canonical connection table representation of the generated constitutional isomers is first processed to identify all stereocenters. A concept of symmetry called the configuration symmetry group is the basis of making the transition from constitutional to configurational isomers. This approach was also used in the extension of the STREC program.[20]

Stereocenter identification also initiates the procedure in the CHEMICS program.[21] Here, all possible combinations of configurational parities are generated in a process that uses defined configurational molds. In a final step, a stereochemically unique name is assigned to each stereoisomer for the purpose of identifying and eliminating redundancies which arise due to the presence of topological symmetry.

A recently described procedure developed for SESAMI starts with the premise that the origin of stereoisomerism in most compounds of interest to chemists is the presence of certain topological features which are defined as stereocenters.[22] The maximum number of stereoisomers possible for a compound with n stereocenters is 2^n. This occurs only when every stereocenter is a “true” stereocenter (e.g., a tetracoordinate carbon atom with four topologically different groups attached) and each is different from the others. This condition may be achieved even in the presence of topological symmetry. However, the presence of topological symmetry in a compound can also lead to topological equivalence among true stereocenters and/or introduction of a second type of stereocenter, designated a para-stereocenter. In these cases, the number of stereoisomers will be less than 2^n.

This approach begins with the identification of all stereocenters and their type (true or para). Initially, every stereocenter is assumed to be a true stereocenter different from every other stereocenter. Since a stereocenter can only have one of two configurations (e.g., *R* or *S*), a compound with n different stereocenters will lead to a set of 2^n different parity combinations (configurations), each of which can be expressed as a binary vector of n elements (stereoparity vector). If the compound actually possesses n different true stereocenters, stereoisomer generation is complete. If, on the other hand, the presence of topologically equivalent true stereocenters and/or para-stereocenters have been identified, the number of stereoisomers will be less than 2^n. In operational terms, that means that redundancies exist in the set of 2^n stereoparity vectors. The objective in this approach is to identify and remove those redundancies. The automorphism group of the compound is the key to translating connectivity to configuration and revealing equivalence among the binary vectors. The output of the program is expressed as a set of augmented connection tables in which the parity of each stereocenter of each stereoisomer is included.

SPECTRUM INTERPRETATION

Introduction. One of the earliest applications of computer-based techniques in decision-making was the reduction of spectral data to their structural implications, information that is usually expressed in terms of substructures predicted to be present or absent in the unknown. From the beginning, the techniques studied generally fall into one of three classes: library search, pattern recognition, or rule-based systems. However, the boundaries between these approaches are not sharp. The output of an interpreter can serve as information to be used by a chemist or spectroscopist, or as input to a structure generator, with or without examination by the user. Substructures predicted to be present in the unknown should have a high confidence level since, if the prediction is invalid, every generated structure will be invalid.

Library Search. In a library search, the source of the structural information about the compound under study is a reference library of spectra of compounds of known structure. The procedures, which involve comparing the observed spectral data of the unknown to the reference spectra of the library, are generally specific to a single spectroscopic method, but, for a given compound, the results from separate searches can be combined to increase both scope and accuracy. The search can be conducted in one of two ways. In each case, a high-quality reference library containing both structures and complete spectra is needed which is broadly representative of the types of compounds encountered in the laboratory.

In a similarity search the goal is to retrieve entries in the library whose spectra are “similar” to that of the unknown. Similarity of spectra is measured in terms of a specific metric. The utility of the method is based on the plausible assumption that the structures corresponding to a set of similar spectra should have similar structural features. In the search, the structures corresponding to the reference spectra retrieved from the database could reveal information about the class of compounds to which the unknown belongs, e.g., steroid, and/or the presence of substructures. The retrieval of more than one compound of the same class and the presence of a substructure common to all or most of the retrieved compounds lead to more secure inferences.

In one approach, a set of features that characterize spectra derived from a given spectroscopic method is defined by knowledgeable chemists and/or spectroscopists. The method requires access to a spectral library representing a diverse array of structures. The spectral features identified form the basis of establishing a characteristic signature for each spectrum which can be expressed as a vector or bit string. Using similarity metrics based on these spectral features, programs can identify spectra with signatures similar to that of the query spectrum.

One of the earliest applications of this approach was to the interpretation of mass spectra. The program STIRS uses a set of 18 spectral features to generate a signature of a mass spectrum.[23] The spectral features selected reflect significant structural information. Using matching metrics to compare signatures, the program identifies a rank-ordered list of the spectra in a large database of mass spectra which are most similar to the query spectrum. The structures corresponding to these spectra are retrieved and examined for common substructures which are then considered to be present in the unknown. In another early application to ^{13}C NMR spectra,[24] each spectrum in the database is first encoded using 79 spectral features. These spectral features, each of which is also selected to reflect some significant structural information, may assume only one of two values, e.g., present or absent. The signature of the unknown—a binary vector of 79 elements—is then compared to the signatures of the reference spectra to produce a rank-ordered list of reference compounds most similar to the unknown. These are then examined by the chemist/spectroscopist for structure class and for common substructures. A similar approach to the interpretation of ^{1}H NMR[25] and to the collective spectral data from multiple sources[26] has been described.

In another early approach, spectra were expressed in terms of a vector, each element of which describes an interval in the spectral range (e.g., a frequency interval in the infrared) or some other related spectral parameter (e.g., Fourier coefficients). Each element is considered to be a coordinate in multidimensional space, and therefore each spectrum is characterized by a point in hyperspace. The similarity between an unknown and a reference entry is measured by the “distance” in hyperspace between the two points. The output is a rank-ordered list of structures corresponding to the spectra with the smallest distance to the query. Similarity searches have been applied to the interpretation of ^{13}C NMR spectra,[27] IR spectra,[28] and MS.[29]

Since each signal of a ^{13}C NMR spectrum reveals information about the chemical environment of a single carbon atom (or atoms belonging to the same symmetry class), correlations between subspectra and substructures can be more distinctive than for other spectroscopic methods. A second type of library search is applicable in this case. The interpretive library search retrieves substructures from a database of assigned ^{13}C NMR spectra (i.e., for each spectrum, complete structural information, the chemical shift, and multiplicity of each signal, and the carbon atom to which it is assigned) which are predicted to be present in the unknown.

Interpretive library searches are of two types: those limited to the retrieval of predefined substructures[30] and those open to the retrieval of any substructure part of the compounds of the library.[31] Interpretive library searches are basically subspectrum matching procedures. Matched signals must be within an established tolerance and are usually required to be of the same multiplicity. The premise implicit in these procedures is that if the spectrum of the unknown and a reference library spectrum have a subspectrum in common, then the corresponding reference substructure is also present in the unknown. In programs that retrieve any substructure contained in the database, the signals of the retrieved subspectrum of the reference entry must correspond to a set of connected carbon atoms, i.e., to a substructure. A more advanced version of such a program estimates the probability that a predicted substructure is indeed present in the unknown.[17]

Pattern Recognition. The extraction of structural information from spectral data was the first extensively studied application of early pattern recognition methods. Two types of techniques have been studied. Supervised methods are limited to one or more predefined structural classes and require representative training sets to develop the classifier. Unsupervised methods partition a set of spectra into clusters with common spectral features. The premise is that clusters with common spectral features will have common structural features as well. No predefined structural classes are required, but the clusters must be examined for the common structural features.

In supervised pattern recognition, much of the early work focused on determining if an unknown belongs to a specific class of compounds, i.e., a two-class problem. The process requires a training set of spectra of compounds both with and without the particular structural feature. Similarities to some library search methods described above will be evident. The performance of a given classifier is sensitive to the way in which the spectral data are represented. Transformations of the spectral data are common, for example, in the case of IR, to Fourier or Hadamard coefficients. The coefficients are expressed as a vector whose elements serve as coordinates in hyperspace. For a satisfactory solution to the two-class problem, the cluster of points representing compounds with and without the structural feature must be sufficiently disjointed to allow a clear distinction between classes to be made. One common approach to class assignment is based on a distance measure; an unknown is assigned to that cluster (class) whose center of gravity is closest to the point in hyperspace occupied by the unknown. Alternatively, it may be possible to identify a decision plane in hyperspace which separates points representing compounds in the different classes. In the case of binary classifiers, performance is described in terms of the probability that a compound predicted to be in a class is actually in that class. Applications to ^{1}H and ^{13}C NMR, IR, and MS have been studied.[32−34]

More recently, neural networks, a form of supervised pattern recognition, have been applied to spectrum interpretation.[35] As with other pattern recognition methods, the relationship between spectral properties and structural features in a molecule need not be specified in advance. The network in effect develops the relationship during the process of training. Early applications addressed the interpretation of data from a single spectroscopic source: IR and MS.[36] One of the major advantages of the neural network is that it is amenable to the simultaneous interpretation of data from several spectroscopic sources. Combinations of IR and ^{13}C NMR have been considered,[37] as well as IR and MS.[38]

Unsupervised methods of pattern recognition have received less attention in spectrum interpretation. In an early application, an ordered binary hierarchical tree was generated from a set of IR spectra of known compounds using the three-distance clustering method.[39] An examination of the tree revealed clusters of leaves (spectra) whose corresponding structures had some structural feature in common. Each of these clusters is defined by a vertex which serves as its “root node”. A query spectrum, which entered at the uppermost root node of the tree and guided through the tree using the three-distance method, will link to the leaf to which it is most similar. In principle, the assignment of a structural feature characteristic of a given cluster can be made as the query passes through its vertex.

Rule-Based Systems. An alternative to procedures requiring spectral libraries, rule-based systems come closest to the process by which the chemist/spectroscopist interprets spectral data. Programs consist of two components: a set of rules which is a representation of the known correlations between spectral and structural features and an interpreter program that accepts spectral data as input and uses the rules and some coded form of “reasoning” to make substructural inferences. The rules are developed by experts in the field from the empirical data available in the primary literature, reference works, and/or databases. Rule-based programs have been referred to as expert systems, although, in general, they fail to perform at the level of the expert in intellectually demanding tasks.

Many of the early implementations of rule-based systems were simply tables correlating spectral and structural features. Such programs were incorporated into some of the earliest comprehensive systems for structure elucidation: CHEMICS,[40] STREC,[41] DENDRAL,[42] and DARC.[43] This approach still finds application in more modern versions of CHEMICS[44] and in the more recently developed program SESAMI.[17] In both CHEMICS and SESAMI the correlations serve to eliminate from consideration fragments from a master list of structural building units.

One of the earliest efforts in rule-based systems utilized known fragmentation and rearrangement processes in mass spectrometry to interpret spectra.[45] This work was coupled with the development of a program, Meta-DENDRAL, designed to interactively aid the chemist in the discovery of new rules governing fragmentation from an examination of the spectra of known compounds.[46]

The program PAIRS is one of the most elaborate early applications of rule-based systems for the interpretation of IR spectral data. In its earliest version, rules for the identification of a set of predefined structural features were hard-coded, which makes their modification difficult.[47] In a later version, the rules were separated from the interpreter so that they could be conveniently altered. A special programming language called CONCISE was designed for this purpose.[48] Input to the program includes the positions of all peaks in the query spectrum and their intensities and widths, the molecular formula, if known, and the sampling method.

The rules are organized in a hierarchical decision tree structure. With the spectrum of the unknown as the query, the interpreter estimates the probability that each of the predefined structural features—most of which are functional groups—is present in the unknown. Many of these features are part of nested sets, each of which represents a major structural class and its subclasses. The decision tree for each class is basically a nested sequence of “if, then, else” statements. At each step, the conditional statements amount to a search of the entered spectral data for compliance with the presence of peaks of specified intensity and width in specified spectral ranges. If the conditions are met, a tentative probability score is assigned which can change with succeeding steps. These probabilities are set by the creator of the rules. An early attempt in the interpretation of mass spectra followed a similar approach.[49]

Two-Dimensional NMR Spectroscopy. The development of 2D NMR for the study of the compounds of carbon gave rise to one of the chemist’s most powerful probes of the carbon backbone of complex structures. Given the nature of the information derived from these experiments—connectivity, often ambiguous, between atoms (C−C, H−H, and C−H) corresponding to signals in 1D spectra—the computer-based interpretation techniques described above are not readily applicable. A later-generation version of the structure generator ASSEMBLE, ASSEMBLE2D, was the first program to include the capability to reduce 2D NMR signal correlations directly to structural information.[50] CHEMICS was also expanded to process 2D NMR data.[51]

Given one-bond C−H correlations derived from an experiment as HMQC, the program ASSEMBLE2D reduces the three-bond H−H correlations derived from the COSY experiment and the one-bond C−C correlations derived from the infrequently used 2D INADEQUATE experiment to carbon atom substructures. In the presence of molecular symmetry, multiple interpretations are possible and the procedure is far more complex. However, ASSEMBLE2D is not capable of interpreting highly ambiguous information such as the HMBC-derived long-range C−H correlations that do not distinguish between two and three (sometimes four) intervening bonds. This limitation was overcome with the development of the structure reduction-based structure generator, COCOA, which is at the core of the SESAMI system, and the 2D NMR interpreter, INFER2D.[52] In this system, the ambiguous, long-range C−H atom correlations are converted to ambiguous C−C atom correlations. They are handed to COCOA which uses them prospectively. The CISOC-CES system also uses structure reduction-based structure generation in the interpretation of 2D NMR data.[53]

EVALUATION OF GENERATED STRUCTURES

Introduction. In general, procedures for spectrum interpretation, whether chemist or computer implemented, rarely extract all of the useful information. Therefore, the information content of the substructural inferences is often insufficient to lead directly to the generation of a single structural assignment. In fact, cases where a great many structures result are not uncommon. This led to the development of computer-based tools to evaluate the candidates produced. The simplest of these are editors which organize the candidates into subsets containing similar structural features. In most such applications, the chemist defines the classifying features to be used which could be, for example, functional groups or skeletal types.[8] The classifying features are usually selected for their ability to aid in distinguishing between the candidates. A fully automated classification procedure has also been described which is based on the combinatorial problem of set covering.[54] The program begins with a small substructure common to all members of the set of generated structures—the “seed”—and produces the user-selected number of subsets each of which is characterized by a mutually exclusive substructure. The program attempts to create equally numbered subsets. Its execution is guided by an information−theoretical criterion.

The symmetry information inherent in the number of signals in the ^{13}C NMR spectrum of a compound relative to the number of its carbon atoms has been used in some early systems to eliminate incompatible structures retrospectively.[55,56] The number of signals expected for each candidate structure is calculated by an automated analysis of its topological symmetry and compared with the number of observed signals. The tool is approximate at best, and caution is required in setting the deviation leading to the elimination of candidate structures for two reasons. First, the structures being examined are constitutional, not configurational, isomers, and the number of distinct signals depends on molecular symmetry, not topological symmetry (e.g., two topologically equivalent carbon atoms may differ stereochemically and therefore magnetically). Second, fortuitous overlaps of signals in the observed spectrum is a possibility.

Spectrum prediction has been a major area of study in distinguishing between the candidate structures. Such approaches are based on comparing the degree of similarity between the predicted spectral properties for each candidate and the observed spectral properties of the unknown using some similarity metric. The degree of similarity is the basis for arranging the candidate structures in order of decreasing probability of being correct.

Applications of spectrum prediction to the evaluation of candidate structures have some special requirements. First, comparisons are to be between predicted and experimental spectra, not relative comparisons between predicted spectra; therefore, the quality of the predicted spectrum of a compound must closely approximate its experimentally determined spectrum. Second, the methods must be applicable to larger, complex, highly-functionalized compounds as well as smaller, simpler ones. Third, spectrum prediction procedures must be sufficiently refined to yield spectral distinctions between isomeric compounds that possess structural similarities which at times can be substantial. Finally, since at times there may be many candidates, the procedures should be computationally efficient.

With few exceptions,[57−59] the information input to most spectrum prediction programs relates to connectivity, not to three-dimensional structure. Spectroscopic methods differ in the extent to which spectral properties are influenced by the stereochemistry of the compound. ^{13}C NMR spectra are quite sensitive; mass spectra are not. Furthermore, depending on the spectroscopic method, variations in sampling can also lead to spectral differences. For these reasons, ranking, rather than excluding, candidates on the basis of spectrum prediction and comparison is preferred.

The most common approaches to predicting spectra are based on substructure−subspectra correlations, linear additivity, rule sets, and empirical modeling. The application of ab initio theory has not proved to be generally useful because three-dimensional coordinates are required and because the theoretical equations necessary to treat real-world structures are complex and can require substantial, if not prohibitive, computation times for solution. Semiempirical quantum chemical methods were briefly considered early on[60] and revisited more recently,[61] but they have not been widely utilized.

With the development of ^{13}C NMR as a practical and powerful structural probe of carbon-based compounds, this spectroscopic method became the focus of numerous studies in spectrum prediction. In contrast to the complex patterns of IR and ^{1}H NMR spectra, ^{13}C NMR data are recorded as sharp signals for each magnetically different carbon atom. The method has the potential to discriminate between structurally similar constitutional isomers because chemical shift is sensitive to the structural environment of each carbon atom.

Recently, some spectrum prediction programs have reached a level of performance in terms of quality and speed to permit the user to place less emphasis on spectrum interpretation to provide sufficient structural information as input to structure generation to limit the plausible candidates to a small number and more emphasis on ranking a larger set structures by spectrum prediction as the method of limiting the number of plausible candidates.

Substructure−Subspectra Correlations. As databases of assigned ^{13}C NMR spectra became available, practical applications of substructure−subspectra correlations in spectrum prediction became feasible.[30,62] A more recent implementation of this approach, program CSEARCH,[63] uses a library of concentrically-layered, carbon-centered fragments of up to five shells which are correlated to the ^{13}C NMR chemical shift of the central atom. For each such fragment, a separate chemical shift range and mean value of the central carbon atom is stored for each of the layered fragments nested within the five-sphere fragment. In program execution, each query structure is first disassembled into component carbon-centered fragments of up to five spheres of nearest neighbors wherever possible. The central carbon atom of each query fragment is then assigned the stored chemical shift value of the corresponding largest-sphered fragment available in the library.

One of the earliest spectrum prediction applications is part of the STREC program. It uses a prepared library of substructure−subspectra correlations to predict ^{1}H NMR, MS, IR, and UV spectral data for each generated structure.[41] Only those candidate structures whose predicted spectral properties differed substantially from the observed data were excluded.

Linear Additivity. The application of additivity rules to the prediction of ^{13}C NMR chemical shifts also received early attention.[64] This approach is based on the observation that, within a particular class of compounds, the contribution of a substituent to the chemical shift of a carbon atom is nearly constant. These contributions—substituent increments—can be determined by analyzing spectra of known compounds. Used together with a set of rules, generally expressed as linear mathematical relationships whose terms are weighted contributions of each substituent up to some predefined bond radius from the central carbon atom, chemical shift predictions can be made. The relationships that have been developed pertain for the most part to specific structural arrangements.

A more recent program, C13Shift,[65] is the first general solution covering a wide variety of carbon atom environments. Initially, each carbon atom in the compound is assigned to one of the substructures of a hierarchical list for which a base chemical shift has been estimated. If an exact match is not found, a base chemical shift is interpolated from a similar substructure that is present in the list. After a carbon atom has been assigned to a substructure, all other atoms in the molecule are treated as substituents and its chemical shift is estimated by drawing on a large library of substituent increments. More recently, the same method has been applied to the prediction of ^{1}H NMR chemical shift.[66]

A procedure described as “optimized prediction of ^{13}C NMR spectra using increments” (OPSI) dynamically generates the required additivity models for an entered compound.[67] The procedure automatically generates the “parent structure” and all other possible substructures of the entered compound. An assigned ^{13}C NMR database is then searched for these substructures, and the additivity models are generated from those that are retrieved. Substituent increments are automatically calculated and used in predicting chemical shifts.

Rule Sets. Since structure generators produce constitutional, not configurational, isomers, mass spectrometry would appear to be especially applicable in spectrum prediction because it is not markedly sensitive to stereochemical differences. Decades of study of the mass spectrometry of organic compounds has led to an extensive understanding of the relationships between structural features and the fragmentation pathways. These can be conveniently expressed in terms of rules. The DENDRAL program was the earliest effort to apply MS prediction to evaluation of candidate structures.[68] On the basis of available knowledge, detailed rules were developed to predict mass spectra. However, the rules were class-specific thereby limiting their applicability to multifunctional compounds.

Knowledge of the distinct pathways of fragmentation is not the only important component of an effective rule-based system. In a complex multifunctional molecule, the presence of a structural feature does not assure that the expected fragmentation pattern will occur; some other process may be more favorable. MASSIMO[69] was developed with this complexity in mind. For a multifunctional compound, the program initially calculates values for electronic and energy parameters expected to influence selection of the initial site of fragmentation. Using these calculated values and a knowledge base containing descriptions of the types of allowed fragmentation and rearrangement pathways, the relative probabilities of all applicable pathways are then evaluated. These probabilities influence whether or not a particular fragmentation pathway will be followed. Pathways for fragmentation of daughter ions are similarly evaluated next. The output is a predicted mass spectrum with detailed peak assignments and a description of the processes leading to them. The program is still under development.

Empirical Modeling. One of the most extensively studied applications of empirical modeling to spectrum prediction focused on ^{13}C NMR spectra.[59] The method seeks to develop a set of linear parametric equations that relate a set of descriptors to the ^{13}C NMR chemical shift of individual atoms in a molecule. The equations take the form

$$S = b_0 + b_1X_1 + b_2X_2 + ... + b_iX_i$$

where S is the predicted chemical shift of a particular carbon atom, X_i is the calculated numerical value for descriptor i, and b_i is the corresponding coefficient which is determined by a multiple linear regression analysis using a dataset of observed chemical shifts taken from a set of known, structurally related compounds. Thus, the equations are structure-class dependent. A predefined set of descriptors encode topological, electronic, and geometric characteristics. (This is one of the few methods of spectrum prediction that considers spatial factors.) For each separate structure class, the complete set of descriptors is screened to eliminate the least influential. The techniques of regression analysis are then used to develop the best set of mathematical models for each structure class that can treat each atom type in that class. In a more recent study, an artificial neural network was shown to be superior to multiple linear regression analysis.[70]

A limitation of the method is that each set of equations is structure-class specific. However, in real-world problems, the compound whose spectrum is to be predicted may not readily fall into a previously studied class. In a novel solution to this problem, a neural network was trained to relate the chemical environment of the carbon atoms in a molecule to the mathematical models developed earlier using specific classes of compounds. The “best” model equation is selected for each query atom by using as input to the trained network a vector describing its chemical environment. The “winning” output neuron identifies the best model.

COMPREHENSIVE STRUCTURE ELUCIDATION SYSTEMS

CHEMICS is one of the earliest comprehensive systems to be developed. The program includes software for the three major components of structure elucidation. In the initial stage, the program compares the IR and the ^{13}C and ^{1}H NMR spectral properties assigned to each member of a set of predefined standard fragments (initially 189 in number; now, in treating a broader range of structures, 630) with those of the collective spectral data of the unknown. Those fragments which are not compatible are discarded. The set from which the discards are made is said to include fragments such that, for the elements covered, there are no limits on the type of structures generated. More recently, interpreted 2D NMR data has been added as a tool in discarding fragments in this initial stage.[71]

The surviving fragments, some of which are invalid, serve as the structural building units in the next step which is structure generation (see STRUCTURE GENERATION). The process is constrained by any user-entered substructural inferences and some of the information inferred from the 2D NMR data. Where applicable, the set of generated structures can be further narrowed on the basis of symmetry considerations.[71] The surviving structures are then ranked in order of decreasing probability of being correct based on the goodness of the fit between observed and predicted ^{13}C NMR, ^{1}H NMR, and MS data.[71] In a recent enhancement, NOE (nuclear Overhauser effect) data have been used to build three-dimensional structures from the two-dimensional candidates.[72]

SESAMI is a more recent development. It places a heavy emphasis on spectrum interpretation to generate a rich pool of structural information in order to achieve its design objective: the direct reduction of the collective spectral properties of an unknown to a single structure or a ranked set of plausible compatible structures which is small in number and exhaustive in scope. As currently reported, SESAMI includes capabilities only in spectrum interpretation and structure generation. The program has been designed to operate either in an automated or interactive mode.

Spectrum interpretation in SESAMI is a two-track procedure. On one track, the molecular formula and the collective spectral properties of the unknown are processed by a program (PRUNE) to give rise to a subset of an exhaustive set of the uniformly-sized, explicitly defined basic units of structure, each of which is an atom-centered fragment with one concentric layer of nearest neighbors (ACF). Although the fragment sets are very different, program execution resembles that used in CHEMICS. Each ACF in the exhaustive set is tested for compatibility with the molecular formula and the observed spectral data. Two-dimensional NMR data, if entered, are also used in pruning fragments from the exhaustive set. Since the ACF is too small a fragment to permit a distinction to be made between each of them based on spectral properties, the set of surviving ACFs will usually contain many more invalid than valid fragments. In fact, the program is biased to retain an invalid fragment rather than risk deleting a valid ACF.

On the second track, another modular program (INTERPRET) accepts the same input of spectral data and produces a set of inferences, substructures predicted to be present or absent in the unknown. These substructures can be of any size, complexity, ambiguity, and degree of overlap with other substructures. One module, INFER2D, is a program for the interpretation a broad range of 2D NMR data useful in structure elucidation work, some of which are expressed as alternative interpretations, e.g., the long-range C−H correlations derived from the HMBC experiment. The output of both spectrum interpretation tracks is handed to the structure reduction-based structure generator COCOA, which exhaustively generates all molecular structures compatible with the input. A recent enhancement to SESAMI provides for the generation of all possible stereoisomers of the constitutional isomers output.[22]

CISOC-SES is another recent program that effectively and efficiently utilizes 2D NMR data. Its function and underlying operation are similar to SESAMI. The program uses 1D and 2D NMR spectral data as input and generates structures in a three-stage, structure reduction-based process.

Several structure elucidation systems, e.g., ACCESS,[73] CSEARCH,[74] EPIOS,[75] and SpecSolv,[76] have tightly linked spectrum interpreters and structure generators. Each utilizes a ^{13}C NMR interpretive library search to infer the presence of explicitly defined substructures contained in a library of such fragments created from a database of assigned ^{13}C NMR spectra. (ACCESS uses spectral data from other sources as well.) In ACCESS, EPIOS, and SpecSolv, the substructures inferred are composed of concentric layers of nearest neighbors about a central atom. ACCESS is based on fragments with a maximum of two layers; EPIOS, one layer; and SpecSolv, two and three layers. CSEARCH uses fragments of three atoms. Their subspectrum matching procedures use a library of correlated substructure−subspectra entries in which each carbon atom is assigned a chemical shift range and signal multiplicity. (SpecSolv alone also uses signal intensity.) Substructures retrieved in this way comprise a "hit list" of structural building units that serves as input to the structure generators of these programs. Some of them are, as expected, invalid.

In these programs, the chemical shift range of the carbon atoms in the fragments of these datasets can be determined either by examining chemical shift distributions, for those fragments for which there exists sufficiently large sets, or, in the case of too few occurrences of the fragment in the database, by spectrum prediction procedures or assignment of some default value.

The structure generators in these systems, although structure assemblers, are more program-specific than general purpose and require most or all of the structural information for structure generation to be provided by built-in spectrum interpreters. Since compatible carbon-centered fragments are retrieved for adjacent carbon atoms, overlap of the fragments is a fact. For example, in EPIOS, which uses single-layered fragments, the central carbon atom of one fragment will be a first-layer atom in another fragment. Such ordered overlap forms the basis of structure generation in these programs.

Structure generation resembles a depth-first tree search. To begin the process, a fragment is selected; in SpecSolv it is the largest fragment with the best match factor, measured in terms of chemical shift difference. Using matched chemical shift pairs (EPIOS) or common atoms or atom sequences (ACCESS, SpecSolv), a second fragment is linked to the first, then a third, etc. In EPIOS and SpecSolv, the process is constrained at each step by predicting chemical shifts for each carbon atom in a partially assembled structure and comparing them with observed spectral data. Partially assembled structures with chemical shift deviations that exceed an established value are discarded or weighted less.

In each of these programs, the substructures available as structural building units are limited to those present in the database used. Output of candidate structures is therefore not exhaustive; it is limited to those structures that can be constructed from the substructure library of each program. Nonetheless, these systems are powerful, in particular in laboratory environments where the classes of compounds studied are limited.

PROBLEM SOLVING: THEN AND NOW

One of the earliest (if not the earliest) recorded examples of the application of computer-based tools to a real-world structure elucidation problem appeared in 1967.[77] The same problem was recently revisited using more recently developed software.[78]

The antibiotic actinobolin was isolated 40 years ago from a fermentation beer. Comparison with an available library of antibiotics suggested that it was unrelated in structure type to compounds known at the time. The starting point for the structure elucidation was a crystalline monoacetate of molecular formula $C_{15}H_{22}N_2O_7$. While not large in comparison to some biomolecules, the compound was still too large and complex to reveal its structure merely by a consideration of its simple chemical properties and the limited spectral data that could be obtained at the time. Conventional practice required degradation of the intact unknown into smaller molecules which were either known or readily characterized.

Since nothing was known about the structure of actinobolin, a series of commonly used degradation reactions were chosen for initial study: oxidation and acid and base hydrolysis. Where an initial degradation product could not be readily characterized, it was subjected to more vigorous reaction conditions or other cleavage reactions. From the collective structural information derived from the degradation studies, the presence of a set of substructures in the intact compound was inferred by the chemist (Figure 1). This set of substructures can be thought of as a partial structure of the unknown.

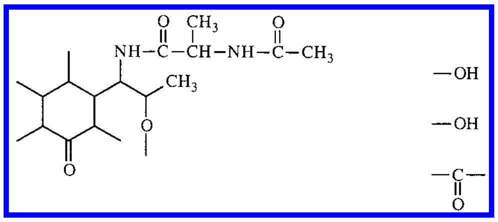

Figure 1. Actinobolin acetate substructures inferred from chemical degradation reactions.

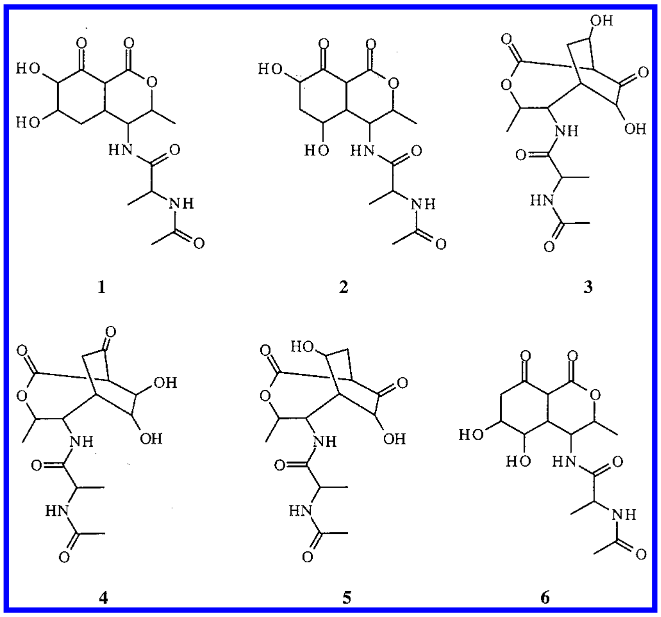

Figure 2. Six candidate structures generated by ASSEMBLE.

A very early version of the structure generator ASSEMBLE was used to expand that partial structure into all molecular structures compatible with it. That structure generation was constrained by the input of some additional information which was deduced by the chemist from the observed chemical behavior of the compound and its spectral data: aldehydes, carboxylic acids, peracids and more than two hydroxyl groups were forbidden; a 1,3-dicarbonyl unit was required. Six structures were produced, all of which were β-ketolactones (Figure 2). Although β-diketones were not explicitly excluded, none were generated. The necessary information is intrinsic in the input.

Since the program provides the assurance that no structure equally compatible with the input has been overlooked, the final assignment of structure was reduced to merely making a distinction between six and only six alternative structures. Examination of the six structures provided invaluable guidance in the design of four experiments which narrowed the choices and led to the assignment of the structure of actinobolin (structure **6**).[77]

In those early years, structure elucidation was largely based on information derived from chemical studies which in the usual case were both tedious and complex. Experimental conditions of the degradation reactions had to be optimized by trial and error. A high order of experimental skill was needed to separate complex reaction mixtures. Often, reaction sequences studied proved to be fruitless. Structure elucidation could be very time-consuming. In the case of actinobolin, several man-years were required, a not uncommon time span for a complex unknown of a type structurally unrelated to known compounds.

The development of SESAMI provided a way of looking at the progress that has been made over the years in computer-enhanced structure elucidation: revisit the actinobolin problem with this more recent software.[78] In contrast to the earlier ASSEMBLE-based study, the only input to SESAMI was 1D and 2D NMR data derived from actinobolin acetate.

Table 1. 1D ^{1}H NMR Spectrum of Actinobolin Acetate

shift	integral	shift	integral	shift	integral
13.18	1 (exch)	4.36	1	2.58	1
7.95	1 (exch)	4.29	1	2.17	1
7.78	1 (exch)	3.54	1	1.77	3
5.01	1 (exch)	3.02	1	1.13	3
4.73	1 (exch)	2.60	1	1.09	3
4.61	1				

Table 2. 1D ^{13}C NMR Spectrum of Actinobolin Acetate

shift	multiplicity	shift	multiplicity	shift	multiplicity
173.97	S	78.08	D	41.17	D
172.99	S	70.73	D	37.09	T
170.64	S	68.46	D	22.32	Q
168.81	S	48.01	D	18.56	Q
91.38	S	44.51	D	17.48	Q

Table 3. HMQC Correlations for Actinobolin Acetate

signal 1[a]	signal 2[a]	min[b]	max[c]
C78.08	H4.61	1	1
C70.73	H3.02	1	1
C68.46	H3.54	1	1
C48.01	H4.36	1	1
C44.51	H4.29	1	1
C41.17	H2.58	1	1
C37.09	H2.60	1	1
C37.09	H2.17	1	1
C22.32	H1.77	1	1
C18.56	H1.13	1	1
C17.48	H1.09	1	1

[a] Element, chemical shift (ppm). [b] Minimun number of intervening bonds. [c] Maximum number of intervening bonds.

Table 4. COSY Correlations for Actinobolin Acetate

signal 1[a]	signal 2[a]	min[b]	max[c]
H7.95	H4.36	3	3
H7.78	H4.29	3	3
H5.01	H3.54	3	3
H4.73	H3.02	3	3
H4.61	H4.29	3	3
H4.61	H1.09	3	3
H4.36	H1.13	3	3
H4.29	H2.58	3	3
H3.54	H2.17	3	3
H3.54	H2.60	3	3
H3.54	H3.02	3	3

[a] Element, chemical shift (ppm). [b] Minimun number of intervening bonds. [c] Maximum number of intervening bonds.

Five of the 16 signals in the ^{1}H NMR spectrum (Table 1) disappeared upon the addition of deuterated water to the sample tube. That hydrogen exchange information is used by the interpretation program of SESAMI together with ^{1}H chemical shift and signal intensity data. Of the available 1D ^{13}C NMR data (Table 2), chemical shift and signal multiplicity due to one-bond carbon−hydrogen coupling are used. The 2D NMR data input were derived from the HMQC experiment (one-bond C−H correlations), the COSY experiment (H−H correlations, usually three-bond), and the HMBC experiment (C−H correlations, usually two- or three-bond, which cannot be distinguished). These correlations are shown in Tables 3−5, respectively. Each entry records the chemical shifts of a pair of correlated atoms and the minimum and maximum number of intervening bonds permitted. With these collective data as input, SESAMI generated four structures in 5 min of CPU time (Figure 3).

Table 5. HMBC Correlations for Actinobolin Acetate

signal 1[a]	signal 2[a]	min[b]	max[c]
C173.97	H1.13	2	3
C173.97	H4.36	2	3
C173.97	H7.78	2	3
C172.99	H2.17	2	3
C172.99	H2.60	2	3
C172.99	H13.18	2	3
C168.81	H1.77	2	3
C168.81	H4.36	2	3
C168.81	H7.95	2	3
C91.38	H2.58	2	3
C91.38	H4.29	2	3
C91.38	H13.18	2	3
C78.08	H1.09	2	3
C70.73	H2.60	2	3
C70.73	H4.73	2	3
C70.73	H5.01	2	3
C68.46	H2.17	2	3
C68.46	H2.58	2	3
C68.46	H4.73	2	3
C68.46	H5.01	2	3
C48.01	H1.13	2	3
C48.01	H7.95	2	3
C44.51	H1.09	2	3
C44.51	H7.78	2	3
C41.17	H4.73	2	3
C37.09	H5.01	2	3
C37.09	H13.18	2	3
C18.56	H7.95	2	3
C18.56	H4.36	2	3
C17.48	H4.61	2	3

[a] Element, chemical shift (ppm). [b] Minimun number of intervening bonds. [c] Maximum number of intervening bonds.

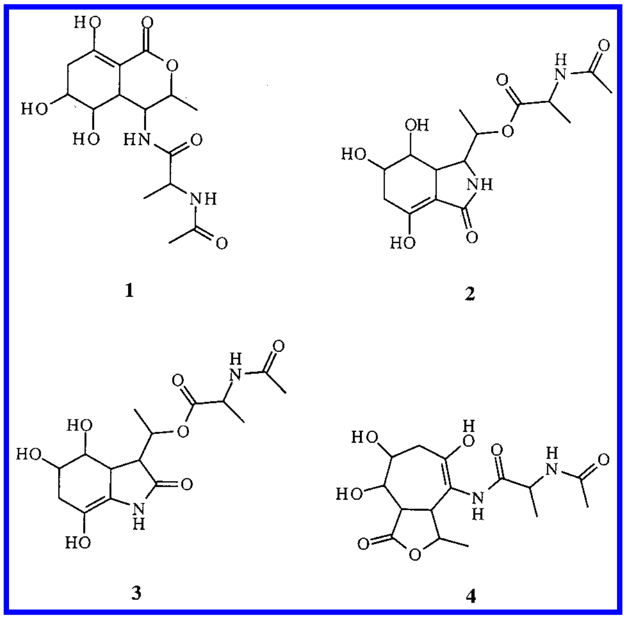

Figure 3. Four candidate structures produced by SESAMI.

Since SESAMI, like ASSEMBLE, is exhaustive, the problem quickly reduced to a distinction between four and only four structures. Two of the structures, **3** and **4**, suffer from chemical instability (the enol/enamine linkage would be expected to exist either as the tautomeric carbonyl compound or the imino compound, neither of which is compatible with the input) and were eliminated from further consideration. As is evident, the remaining two structures are very similar; thus, the choice was not simple. However, a careful examination of the structures suggested that subtle differences in the ^{1}H NMR spectra—specifically, in coupling constants—should be observable. This led to the assignment of structure **1**, consistent with the earlier assignment.[78]

Forty years ago, the actinobolin problem required several man-years to solve and was heavily based on chemical behavior. Now, with more sophisticated spectroscopic methods and computer software, the time required has been reduced to, at most, several days (if data collection time is included), an increase in productivity by at least 2 orders of magnitude. The more sophisticated spectroscopic methods are a major factor in productivity enhancement, but the role of the software should not be underestimated. In particular, the degree of ambiguity in the 30 long-range C−H correlations (Table 5), each of which includes one valid and one invalid correlation, is enormous. (That set of 30 either/or constraints would give rise to over 1 billion (2^{30}) different sets of 30 unambiguous constraints.) However, the program uses that information prospectively and efficiently without any preprocessing. It would be tedious at best for a chemist to process such ambiguous information prospectively.

SUMMARY AND CONCLUSIONS

Each of the three major components of structure elucidation—spectrum interpretation, structure generation, and spectrum prediction—have been shown to be amenable to computer modeling. A variety of promising approaches have been examined in each case. Although early efforts provided more of a foundation for future work rather than useful laboratory tools, more recently, computer programs of considerable practical value to practicing chemists and spectroscopists have been developed and are beginning to reach the marketplace. The availability of spectroscopic databases of increasing diversity and quality has contributed and will continue to contribute to this development. Given the current state-of-the-art in this area and the increasing level of awareness of its importance to practicing chemists, it is realistic to expect far more powerful computer-based tools for structure determination in the near future.

This review is not intended to be comprehensive but rather to illustrate the concepts with selected examples. In any comprehensive description, the work of many other investigators in this field should receive equal attention.[2,79,80]

REFERENCES AND NOTES

(1) 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.

(2) Munk, M. E.; Madison, M. S. In *Encyclopedia of Computational Chemistry;* Schleyer, P. v. R., Ed.; John Wiley and Sons: New York, in press.

(3) Badertscher, M.; Bischofberger, K.; Pretsch, E. *Trends Anal. Chem.* **1980**, *16*, 234−241.

(4) Shelley, C. A.; Hays, T. R.; Munk, M. E.; Roman, R. V. *Anal. Chim. Acta* **1978**, *103*, 121−132.

(5) Masinter, L. M.; Sridharan, N. S.; Lederberg, J.; Smith, D. H. *J. Am. Chem. Soc.* **1974**, *96*, 7702−7714.

(6) Shelley, C. A.; Munk, M. E.; Roman, R. V. *Anal. Chim. Acta* **1978**, *103*, 245−251.

(7) Carhart, R. E. *Technical Report MIP-R-118*; Machine Intelligence Research Unit, University of Edinburgh: Edinburgh, Scotland, 1977.

(8) Shelley, C. A.; Munk, M. E. *Anal Chim. Acta* **1981**, *133*, 507−516.

(9) Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C. *J. Org. Chem.* **1981**, *46*, 1708−1718.

(10) Sasaki, S.; Abe, H.; Ouki, T.; Sakamoto, M.; Ochiai, S. *Anal. Chem.* **1968**, *40*, 2220−2223.

(11) Kudo, Y.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 43−56.

(12) Gribov, L. A.; Elyashberg, M. E. *J. Mol. Struct.* **1970**, *5*, 179−204.

(13) Serov, V. V.; Elyashberg, M. E.; Gribov, L. A. *J. Mol. Struct*. **1976**, *31*, 381−397.

(14) Debska, B.; Duliban, J.; Guzowska-Swider, B.; Hippe, Z. *Anal Chim. Acta* **1981**, *133*, 303−318.

(15) Kleywegt, G. J.; Luinge, H. J.; van't Klooster, H. A. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 291−302.

(16) Christie, B. D.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87−93.

(17) Munk, M. E.; Velu, V. K.; Madison, M. S.; Robb, E. W.; Baderstscher, M.; Christie, B. D.; Razinger, M. In *Recent Advances in Chemical Information II*; Collier, H., Ed.; Royal Society of Chemistry: Cambridge, U.K., 1993; pp 247−263.

(18) Nourse, J. G.; Carhart, R. E.; Smith, D. H.; Djerassi, C. *J. Am. Chem. Soc.* **1979**, *101*, 1216−1223.

(19) Nourse, J. G.; Smith, D. H.; Carhart, R. E.; Djerassi, C. *J. Am. Chem. Soc.* **1980**, *102*, 6289−6295.

(20) Zlatina, L. A.; Elyashberg, M. E. *MATCH* **1992**, *No. 27*, 191−207.

(21) Abe, H.; Hayasaka, H.; Miyashita, Y.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 216−219.

(22) Razinger, M.; Balasubramanian, K.; Perdih, M.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 812−825.

(23) Kwok, K.-S.; Venkataraghavan, R.; McLafferty, F. W. *J. Am. Chem. Soc.* **1973**, *95*, 4185−4210.

(24) Schwarzenbach, R.; Meilli, J.; Könitzer, H.; Clerc, J.-T. *Org. Magn. Reson.* **1976**, *8*, 11−16.

(25) Farkas, M.; Bendl, J.; Welti, D. H.; Pretsch, E.; Dütsch, S.; Portmann, P.; Zürcher, M.; Clerc, J.-T. *Anal. Chim. Acta* **1988**, *206*, 173−187.

(26) Naegeli, P. R.; Clerc, J.-T. *Anal. Chem.* **1974**, *45*, 739A−744A.

(27) Zupan, J.; Penca, M.; Hadzi, D.; Marsel, J. *Anal. Chem.* **1977**, *49*, 2141−2146.

(28) Kowalski, B. R.; Jurs, P. C.; Isenhour, T. L.; Reilley, C. N. *Anal. Chem.* **1969**, *41*, 1945−1949.

(29) Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L. *Anal. Chem.* **1969**, *41*, 21−27.

(30) Bremser, W.; Klier, M.; Meyer, E. *Org. Magn. Reson.* **1975**, *7*, 97−106.

(31) Shelley, C. A.; Munk, M. E. *Anal. Chem.* **1982**, *54*, 516−521.

(32) Varmuza, K. *Pattern Recognition in Chemistry*; John Wiley and Sons: New York, 1980.

(33) Luinge, H. J. In *Computing Applications in Molecular Spectroscopy*; George, W. O., Steele, D., Eds.; Royal Society of Chemistry: Cambridge, U.K., 1995; pp 87−103.

(34) Woodruff, H. B.; Snelling, C. R.; Shelley, C. A.; Munk, M. E. *Anal. Chem.* **1977**, *49*, 2075−2080.

(35) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH Publishers: New York, 1993.

(36) Citations numbered 5−15 in: Munk, M. E.; Madison, M. S.; Robb, E. W. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 231−238.

(37) Munk, M. E.; Madison, M. S.; Robb, E. W. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 231−238.

(38) Klawun, C.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 249−257.

(39) Zupan, J.; Munk, M. *Anal. Chem.* **1985**, *57*, 1609−1616.

(40) Sasaki, S.; Kudo, Y.; Ochiai, S.; Abe, H. *Mikrochim. Acta* **1971**, 726−742.

(41) Gribov, L. A.; Elyashberg, M. E.; Serov, V. V. *Anal. Chim. Acta* **1977**, *95*, 75−96.

(42) Mitchell, T. M.; Schwenzer, G. M. *Org. Magn. Reson.* **1978**, *11*, 378−384.

(43) Dubois, J.-E.; Carabedian, M.; Ancian, B. *C. R. Acad. Sci. (Paris)* **1980**, *290*, 369−372, 383−386.

(44) Funatsu, K.; Miyabayashi, N.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18−28.

(45) Smith, D. H.; Buchanan, B. G.; Engelmore, R. S.; Duffield, A. M.; Yeo, A.; Feigenbaum, E. A.; Lederberg, J.; Djerassi, C. *J. Am. Chem. Soc.* **1972**, *94*, 5962−5971.

(46) Buchanan, B. G.; Smith, D. S.; White, W. C.; Gritter, R. J.; Feigenbaum, E. A.; Lederberg, J.; Djerassi, C. *J. Am. Chem. Soc.* **1976**, *98*, 6168−6178.

(47) Woodruff, H. B.; Munk, M. E. *J. Org. Chem.* **1977**, *42*, 1761−1767.

(48) Woodruff, H. B.; Smith, G. M. *Anal. Chem.* **1980**, *52*, 2321−2327.

(49) Gray, N. A. B.; Gronneberg, T. O. *Anal. Chem.* **1975**, *47*, 419−424.

(50) Christie, B. D.; Munk, M. E. *Anal. Chim. Acta* **1987**, *200*, 347−361.

(51) Funatsu, K.; Susuta, Y.; Sasaki, S. *J. Chem Inf. Comput. Sci.* **1989**, *29*, 6−11.

(52) Christie, B. D.; Munk, M. E. *J. Am. Chem. Soc.* **1991**, *113*, 3750−3757.

(53) Peng, C.; Yuan, S.; Zheng, C.; Hui, Y. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 805−813.

(54) Lipkus, A. L.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 9−18.

(55) Shelley, C. A.; Munk, M. E. *Anal. Chem.* **1978**, *50*, 1522−1527.

(56) Fujiwara, I.; Okujama, T.; Yamaski, T.; Abe, H.; Sasaki, S. *Anal. Chim. Acta* **1981**, *133*, 527−533.

(57) Gray, N. A. B.; Crandell, C. W.; Nourse, J. G.; Smith, D. H.; Dageforde, M. L.; Djerassi, C. *J. Org. Chem.* **1981**, *46*, 703−715.

(58) Gray, N. A. B.; Nourse, J. G.; Crandell, C. W.; Smith, D. H.; Djerassi, C. *Org. Magn. Reson.* **1981**, *15*, 375−389.

(59) Jurs, P. C.; Ball, J. W.; Anker, L. S.; Friedman, T. L. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 272−278.

(60) Gribov, L. A.; Elyashberg, M. E.; Raikhshtat, M. M. *J. Mol. Struct.* **1979**, *53*, 81−94.

(61) Tusar, M.; Tusar, L.; Bohanec, S.; Zupan, J. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 299−303.

(62) Bremser, W. *Z. Anal. Chem.* **1977**, *286*, 1−13.

(63) Kalchhauser, H.; Robien, W. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 103−108.

(64) Clerc, J. T.; Sommerauer, H. *Anal. Chim. Acta* **1977**, *95*, 33−40.

(65) Pretsch, E.; Furst, A.; Badertscher, M.; Burgin, R.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 291−295.

(66) Burgin-Schaller, R.; Junghans, M.; Schriber, H.; Badertscher, M.; Pretsch, E.; Munk, M. E. In *Software Development in Chemistry*; Moll, R., Ed.; Springer Verlag: Berlin, 1995; Vol. 9, pp 241−256.

(67) Chen, L.; Robien, W. *Anal. Chem.* **1993**, *65*, 2282−2287.

(68) Schroll, G.; Duffield, A. M.; Djerassi, C.; Buchanan, B. G.; Sutherland, G. L.; Feigenbaum, E. A.; Lederberg, J. *J Am. Chem. Soc.* **1969**, *91*, 7440−7443.

(69) Gasteiger, J.; Hanebeck, W.; Schulz, K.-P.; Bauerschmidt, S.; Hollering, R. In *Computer-Enhanced Analytical Spectroscopy*, Vol. 4; Wilkins, C. L., Ed.; Plenum Press: New York, 1993; pp 97−133.

(70) Jurs, P. C.; Anker, L. S.; Ball, J. W. In *Computer-Enhanced Analytical Spectroscopy*, Vol. 4; Wilkins, C. L., Ed.; Plenum Press: New York, 1993; pp 1−35.

(71) Funatsu, K.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 190−204.

(72) Funatsu, K.; Nishizaki, M.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 745−751.

(73) Bremser, W.; Fachinger, W. *Magn. Reson. Chem.* **1985**, *23*, 1056−1071.

(74) Robien, W. *Mikrochim. Acta* [Wien] **1986**, *II*, 271−279.

(75) Carabedian, M.; Dagane, I.; Dubois, J.-E. *Anal. Chem.* **1988**, *60*, 2186−2192.

(76) Will, M.; Fachinger, W.; Richert, J. R. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221−227.

(77) Munk, M. E.; Sodano, C. S.; McLean, R. L.; Haskell, T. H. *J. Am. Chem. Soc.* **1967**, *90*, 4158−4165.

(78) Madison, M. S.; Schulz, K.-P.; Korytko, A. A.; Munk, M. E. *Internet J. Chem.*, in press.

(79) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; John Wiley and Sons: New York, 1986.

(80) Hippe, Z. *Artificial Intelligence in Chemistry-Structure Elucidation and Simulation of Organic Reactions*; Elsevier: New York, 1991.

CI980083R