SENECA: A Platform-Independent, Distributed, and Parallel System for Computer-Assisted Structure Elucidation in Organic Chemistry

Christoph Steinbeck†

Max-Planck-Institut für Chemische Ökologie, Carl-Zeiss-Promenade 10, D-07745 Jena, Germany

Received December 27, 2000

The program package SENECA for Computer-Assisted Structure Elucidation (CASE) of organic molecules is described. SENECA is written completely in the programming language Java and divided into a server, a client, and a gatekeeper part. While the client allows for input of spectroscopic information, the server part performs the actual structure elucidation by stochastically walking through constitution space while optimizing the molecule toward agreement with given spectral properties. The convergence is guided by simulated annealing. The gatekeeper administers a list of server processes, which can be retrieved by the client. The package is completely platform-independent and its server part can be distributed over the Internet or an intranet using a heterogeneous network of almost any number and type of computers, thus allowing for parallel CASE computations on ordinary networks, present in almost any institution.

INTRODUCTION

Structure elucidation and identification is an integral part of organic chemistry. Synthetic chemists need to characterize their reaction products and interesting byproducts, while chemists working on the isolation and structure elucidation of natural products initially do not even know to which class their isolated compound belongs. The vast number of compounds produced by combinatorial chemistry techniques introduces a new dimension to the structure elucidation problem. Spectroscopic methods, such as mass spectrometry (MS), nuclear magnetic resonance (NMR), and IR and UV/VIS spectroscopy have, thus, become indispensable tools in chemistry. This recognition has led to substantial effort, beginning about 20 years ago, to automate the structure elucidation process.[1]

With the arrival of 2D NMR spectroscopy in the 1970s and its development into a routine method in chemical laboratories, it has become clear that many CASE problems can be solved using the molecular formula information extracted from high resolution mass spectra and 2D NMR connectivity data. A signal in a 1D proton NMR spectrum, at a particular frequency, can arise from a number of different proton types, in a number of different chemical environments. A cross signal in a 2D NMR spectrum clearly links the two nuclei involved in a yes/no fashion, depending on the kind of experiment that is performed. As opposed to the "fuzzy" or "analog" information in many 1D NMR spectra one could call this the "digital" type of spectroscopic information. This "digital" information is not always unambiguous. In the case of HMBC spectra, for instance, a cross signal tells the spectroscopist that the proton resonating at frequency f1 is either separated by two, three, or sometimes four bonds from a carbon atom resonating at frequency f2 (so-called $^2J_{CH}$, $^3J_{CH}$, and $^4J_{CH}$ couplings). I have used this information in an earlier work, by designing the program LUCY.[2] Herein, Computer Assisted Structure Elucidation (CASE) was performed, based on the knowledge of the molecular composition of the compound in question and the standard suite of NMR experiments (1D NMR: ^{13}C NMR, DEPT-90, DEPT-135; 2D NMR: HH-COSY, HSQC, HMBC), with particular emphasis on the ambiguous HMBC information described above. LUCY was able to determine the structure of unknown compounds with up to around 30 heavy (non-hydrogen) atoms (see also bracketed information in the following paragraph), provided that it was supplied with absolutely error free information, extracted from the NMR spectra by the spectroscopist. This was achieved by exhaustive generation of all structures that were compatible with the given input data, while using the ambiguous HMBC information in a preprocessing step, prior to the actual structure generation process. This led to a substantial reduction in calculation time and made possible the CASE of molecules up to the size limit mentioned above. A number of programs have arisen since then that use essentially the same approach (for recent reviews[1]).

Although successfully performing CASE on structures that already had "real life" sizes, LUCY had limitations that were inherent to its design. First, the indispensable preprocessing step, using the HMBC information, made use of the assumption that a HMBC signal can only arise from $^2J_{CH}$ and $^3J_{CH}$ couplings, ignoring the fact that there is a small but significant set of signals caused by $^4J_{CH}$ coupling in the spectra. Spectral data containing $^4J_{CH}$-peaks (indistinguishable from the other types) thus led to both false-positive and false-negative results. The second limitation was due to the principle of exhaustive generation of all constitutions in agreement with the input data. The current limit in the size of treatable molecules cannot be extended much further, due to the combinatorial explosion when going to larger numbers of heavy atoms. To date, none of the CASE programs using deterministic structure generation have significantly exceeded

† Corresponding author phone: +49 (0) 3641-643644; fax: +49 (0) 3641-643665; e-mail: steinbeck@ice.mpg.de.

10.1021/ci000407n CCC: $20.00
Published on Web 10/05/2001

the 30-heavy-atoms limit. [There are examples where larger molecules could be treated, based on exceptionally well-defined data sets. Here, large pieces of the molecular skeleton can be defined by synergistically using HSQC and HH-COSY or even 1,1-adequate information before even invoking the structure generator. I do not consider these examples to be valid knock-down arguments for my statement.] SENECA, as described in this article, is designed to address these issues and to overcome LUCY's limitations.

MATERIALS AND METHODS

Simulated Annealing. Recently, a stochastic algorithm has been suggested by Faulon to overcome the limitation of CASE to relatively small molecules.[3] The algorithm's ability to optimize even the constitution of larger molecules toward having the highest possible Wiener Index by Simulated Annealing (SA) was shown.

A significant number of optimization problems in science cannot be tackled with deterministic algorithms simply because the search space is too large to be searched completely within a reasonable time span. Stochastic methods have thus been developed that are able to find the optimum or a near-optimal solution. One of them is Simulated Annealing, another is the class of Genetic Algorithms. Optimization with Simulated Annealing was first suggested by Kirkpatrick et al. They "show, how the Metropolis algorithm"[4] "for approximate numerical simulation of the behavior of a many body system at a finite temperature provides a natural tool for bringing the techniques of statistical mechanics to bear on optimization". Within our frame of reference—the space of constitutional isomers in accordance with a particular molecular formula—this can be described as follows. Based on the input of molecular formula (MF) of the unknown molecule, derived for example from high-resolution mass spectra, and a set of constraints derived from NMR and other spectroscopic data, the process starts with a randomly generated structure S1. Energy E1 is calculated which is a measure of the agreement of this structure with the given (spectroscopic) constraints. The structure is then randomly transformed (see below) into another isomer S2 of the given MF and a new energy value E2 is determined for the newly generated structure. If E2 < E1 (the new structure has a better agreement with the spectroscopic data than the former one) that structure is accepted. If E2 > E1, a random number R with $0 < R < 1$ is generated and the structure is accepted if

$$R < e^{-(E2-E1/kT)} \tag{1}$$

This simulation process, when run at a constant, finite temperature T, converges to the Boltzmann distribution. Kirkpatrick et al. showed that, when running the Metropolis-simulations at successively lower temperatures, only the optimum would be sufficiently populated at sufficiently low temperatures and the systems "freezes" into an optimum state.[5] In SENECA I employ this principle, using an annealing schedule as described below, to determine the constitution of an unknown compound based on an open set of spectroscopic methods, including 1D BB-decoupled carbon NMR, DEPT, HH COSY, HSQC, HC HMBC, HN HMBC. In other words, SENECA attempts to optimize a multivariate, nonlinear, discrete, and constrained target function. While these methods mentioned are solely NMR experiments, it will become clear that the system is by no means limited to a particular type of spectroscopy.

The Structure Generator. Faulon provides a set of simple equations based on which a structure can be randomly converted in another structure, in agreement with chemical rules and the molecular formula.[3] Let x_1, y_1, x_2, and y_2 be the four selected atoms. Let a_{11}, a_{12}, a_{21}, and a_{22} be the orders of the bonds $[x_1,y_1]$, $[x_1,y_2]$, $[x_2,y_1]$, and $[x_2,y_2]$ in the initial molecular graph, and let b_{11}, b_{12}, b_{21}, and b_{22} be the order of the same bonds after the SA random displacement is achieved.

$$b_{11} \geq MAX(0, a_{11} - a_{22}, a_{11} + a_{12} - 3, a_{11} + a_{21} - 3) \tag{2}$$

$$b_{11} \leq MIN(3, a_{11} + a_{12}, a_{11} + a_{21}, a_{11} - a_{22} + 3) \tag{3}$$

$$b_{12} = a_{11} + a_{12} - b_{11} \tag{4}$$

$$b_{21} = a_{11} + a_{21} - b_{11} \tag{5}$$

$$b_{12} = a_{22} - a_{11} + b_{11} \tag{6}$$

To calculate the random displacement in constitution space, choose a random value for $b_{11} \neq a_{11}$ and verify eqs 2 and 3. The values for b_{12}, b_{21}, and b_{22} are calculated using eqs 4−6.

The Annealing Schedule. The calculation is steered toward an optimum by an annealing schedule that describes how the temperature changes in time, i.e., depending on the numbers of displacement steps that have been performed. This section outlines a standard way of performing the cooling which is also the standard implementation of the AnnealingSchedule class in SENECA.

Typically, the temperature T_k at a given step k is calculated as a fraction $T_k = \alpha T_{k-1}$ of the temperature T_{k-1} at a previous step $k-1$. Within such an exponential cooling schedule, cooling rates $\alpha \approx 0.95$ are recommended.[5] At each temperature a sequence of calculations is run, as if to reach a thermal equilibrium. The question of how many equilibrating steps to run at each temperature seems to be one of the key questions of simulated annealing. Several solutions have been proposed. I have adopted one which stops the equilibration once a configurable number of uphill moves have been accepted, or a total number of steps per temperature plateau has been reached, whatever comes first. The latter criterion accounts for the fact that the number of accepted uphill moves becomes very small as the temperature approaches zero.

Another question is the choice of the starting temperature T_0. It must be sufficiently high to ensure that any point of the search space can be reached at the beginning of the calculation but as low as possible for not wasting too much time in inefficient cooling. Instead of choosing an arbitrary "high" value, it is common practice to choose the starting temperature such that approximately 80% of the uphill moves are accepted. Lower values have been proposed.[6] T_0 is estimated by conducting an initial search in which all uphill moves of the target function f are accepted and calculating

the average objective increase $\delta^- f^+$ observed. For a given initial acceptance ratio χ_0, T_0 can then be calculated by $T_0 = (\delta^- f^+)/\ln(\chi_0)$ because χ_0 is not introduced.

Since good annealing schedules are as much subject of ongoing research as they seem to be somewhat of a black art, SENECA has a plug-in architecture that allows contributors to add their own AnnealingSchedule implementations. A default implementation is given.

The Energy Function. Introduction. After each random displacement in constitution space, the newly generated structure is evaluated with respect to its agreement with the given spectroscopic properties. For most types of spectra there is a single type of "Judge" that assigns points to the structures. The points sum up to yield an overall score (eq 7)

$$E_{tot} = E_{HBMC} + E_{HHCOSY} + E_{Shift} + E_{Symmetry} + ... + E_{Features} \quad (7)$$

In the following the functionality of each Judge is briefly described.

The TwoDSpectrumJudge. This is the base class for the evaluation of 2D spectra correlating signals of nuclei by *J*-couplings via multiple bonds, as in HC HMBC, HN HMBC, HH COSY, and others. It is inherited by specialized classes such as HMBCJudge and HHCOSYJudge. They operate on lists of pairs of heavy atoms, so-called TwoDRules, that are somehow (depending on the spectrum type they arise from) correlated by the underlying 2D NMR spectral data. The process for producing these lists is outlined below for the HMBC example. First, a list of heavy atoms is produced based on the molecular formula entered by the user. The signals in 1D carbon NMR spectrum are sequentially assigned to each of the carbon atoms. If there are degenerated resonances, they are assigned to multiple carbon atoms. This step may require user intervention. The DEPT spectra do not allow for the assignment of the number of hydrogen atoms to all carbon atoms. Surplus hydrogens are assigned to other heavy atoms. Ambiguities arising during this step must be resolved by the user. Also considering the HSQC spectra each of the carbon atoms is labeled with a carbon chemical shift and, if it carries hydrogen atoms, also with a proton chemical shift. For each of the two chemical shifts found in an HMBC cross signal the corresponding heavy atom is searched, and a TwoDRule is formed for the atom pair. Due to a degeneracy of the carbon spectrum this assignment process might not be unambiguous. If a particular chemical shift fits multiple heavy atoms, a TwoDRule is created for each possible pair.

When an instance of the TwoDSpectrumJudge is asked to evaluate a structure it starts a Single Source Shortest Path (SSSP) search from one atom of each TwoDRule until it either finds the other atom in the rule or until a cutoff path length is reached (usually shorter than four edges). A number of points $P(|p_{i,j}|)$ are granted, as defined by the user, if a certain path length $|p_{i,j}|$ between two heavy atoms atoms i and j is given, which makes the observed cross-peaks in HMBC or COSY spectra possible (eq 8).

$$P_{tot} = \sum_{crosspeaks} P(|p_{i,j}|) \quad (8)$$

This principle allows for tolerance against unusual but possible cross-peak types such as $^4J_{CH}$ couplings in HMBC spectra. The HMBCJudge can thus be configured, for example, to grant 100 points for each satisfied $^2J_{CH}$ and $^3J_{CH}$ coupling and five points for $^4J_{CH}$ couplings. In a similar manner, the HHCOSYJudge can grant points not only for satisfied $^3J_{HH}$ couplings but also a smaller number of points for less likely couplings over longer ranges.

The HOSECodeJudge. If the structure elucidation is solely guided by TwoDSpectrumJudges, the solutions generated by the structure generator often tend to be highly bridged, and the atoms will not have reasonable atom environments in terms of hybridization and heteroatom attachments. [Of course this is only true in the case of unsaturated compounds where the double bond equivalents can also be satisfied by forming additional rings.] While there initially were (and still are available, if desired) a HybridizationJudge and a HeteroAtomJudge that grant points if an inspected atom has the correct hybridization [as assigned by the user upon inspection of the ^{13}C spectrum] or is or is not attached to a heteroatom, respectively, I was looking for a more automated, less error-prone method. For a human expert the carbon chemical shift is a rich source of information regarding hybridization and heteroatom attachments, so I decided to employ a simple back-calculation of carbon chemical shifts and a comparison with the experimental values to guide the atom environments into the correct state during Simulated Annealing. The only method that is currently fast enough for the inspection of a couple of hundred or thousand structures per second is the shift calculation based on HOSE codes,[7] which describe the spherical environments of a carbon atom. This method requires tables with expectation values for the carbon chemical shift of a carbon atom with a given HOSE code. These tables are compiled from large NMR spectral databases, like the SpecInfo database (http://www.fiz-karlsruhe.de/stn/Databases/specinfo.html). To achieve a reasonable prediction quality in spectra prediction, four- and five-sphere HOSE codes are needed. One can easily imagine the large number of possible HOSE codes in organic molecules, even if only four spheres are involved. Given the immense effort and manpower to assemble the large shift databases needed to create HOSE code expectation values, it becomes clear that the available shift prediction tools are all commercial products. For a free program like SENECA it was not possible to use one of these commercial tools so I had to look for other solutions.

It turned out that for the purpose of generating correct atoms environments, during the stochastic structure generation, the prediction quality of even one-sphere HOSE codes was sufficient, although using more spheres would certainly be an interesting option. Fortunately, there was also at least one reasonably complete table of 651 one-sphere HOSE codes published by Bremser,[8] based on a collection of about 58 000 ^{13}C NMR spectra. For each code, Bremser gave the mean value of all chemical shifts observed for a given HOSE code (δ_{pred}), the 95% confidence limit (CL) for each shift, based on the number of lines the mean value was calculated from, as well as the lowest (δ_{min}) and the highest (δ_{max}) observed shift for this kind of environment. This table forms the basis on which SENECA's HOSECodeJudge grants a sum of points p to a structure according to eq 9.

$$p = \sum_{carbons} \begin{cases} p_{full} & \text{if } \delta_{pred} - CL < \delta_{obs} < \delta_{pred} + CL \\ p_{full}\left(1 - \dfrac{|\delta_{pred} - \delta_{obs}|}{|\delta_{pred} - \delta_{\min}|}\right) & \text{if } \delta_{\min} < \delta_{obs} < \delta_{pred} - CL \\ p_{full}\left(1 - \dfrac{|\delta_{pred} - \delta_{obs}|}{|\delta_{pred} - \delta_{\max}|}\right) & \text{if } \delta_{\max} > \delta_{obs} > \delta_{pred} + CL \\ 0 & \text{if } \delta_{\max} < \delta_{obs} \text{ or } \delta_{\min} > \delta_{obs} \end{cases} \quad (9)$$

The SymmetryJudge. If there is evidence for symmetry in the molecule, for example because there is a lower number of carbon signals in the spectrum than one would expect from the molecular formula, the SymmetryJudge can be used to favor structures. Carbon atoms with the same carbon chemical shift may thus be assigned to the same symmetry class *S*. During the structure generation, Morgan's extended connectivity values[9] EC_i are calculated for each atom i in the candidate compound. If all atoms in *S* have the same EC value, a predefined number of points p_{full} is granted (eq 10).

$$p = \sum_{\text{symmetry classes S}} \begin{cases} p_{full} & \text{if } EC_i = EC_j \text{ for all pairs of atoms } i,j \in S \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

While this is not a rigorous way of symmetry perception, it is very fast, and no symmetry information is missed because the algorithm only fails with false positive results. Methods for computing vertex invariants with higher discriminative power have recently been published by Faulon[10] as well as Xu and Johnson.[11] They might be considered for implementation in future releases.

The FeatureJudge. It is sometimes desirable to be able to determine that certain features may be present or absent in solution structures, i.e., to forbid bonds between heteroatoms or restrict the rings that may be present to certain sizes. The FeatureJudge is a container for these features. For ring size restriction, for example, Figueras' fast algorithm for the perception of the Smallest Set of Smallest Rings[12] is used, and penalties are given if forbidden rings of a certain size are present. This is a good example for the advantages of a scoring system. The score may be chosen high enough such that the existence of a certain ring size is a knockout criterion. But, it may also be chosen low enough such that the ring penalty can be overruled by other overwhelming evidence, for the existence of such a ring, from the rest of the spectroscopic information.

Distributed Computing. If the parameters are carefully selected, a single simulated annealing run has a high probability of finding the correct global optimum. However, since it is a stochastic process, there is no guarantee of this. It is thus common practice to run the process multiple times and to compare the results. [The 3D solution structures of biomolecules are often determined by restrained molecular dynamics based on NMR NOE data. A simulated annealing algorithm drives the convergence of these dynamics runs toward the correct structure. The reader might remember the typical least-squares fit overlay of an ensemble of the best structures as wire-frame models in publications from this field. This is beyond other reasons done to show sufficient convergence of the structure determination to the same overall structural motif.] If a majority of the processes come to the same result, the chances are high that it is the optimum. If not, one should modify the annealing schedule toward slower cooling and/or modify the scoring functions.

The calculations employed here are completely independent of each other, so that the process is highly parallelizable. However, since the purpose of running it in parallel is the statistical verification of the optimum, there is no point in going beyond a certain number of nodes, say 16, for the parallel computation.

SENECA's server part is designed to be distributed over a potentially large number of computers of any type able to run Java programs. The so-called Remote Method Invocation (RMI), an integral part of the Java core API, allows for easy communication of Java programs over networks and brings the distributed architecture, sketched below, easy to hand.

The SENECA Server. The StochasticGeneratorServer is a small compute engine running in the background of either a dedicated machine, for example in a compute cluster, or of a desktop computer also used for regular office work. When contacted by a client, it instantiates a StochasticGenerator object and retrieves a set of Judges and an AnnealingSchedule object via the network connection. It then independently performs the simulated annealing optimization, while providing methods for client to observe the progress of this process. The client contacts the server again upon completion of the calculation to retrieve the results.

The SENECA Gatekeeper. Inspired by other experiments [like the Great Internet Mersenne Prime Search (http://www.gimps.org), the Search for Extraterrestrial Intelligence SETI@home (http://setiathome.ssl.berkeley.edu), Folding@home (http://foldingathome.stanford.edu), or Distributed.net (http://distributed.net)] that harvest the processing power of thousands of desktop machines[13] around the world I added a "Gatekeeper" part to SENECA which coordinates the computing effort. While, in an intranet, a subnet scan can be used to find all the machines on which a SENECA server is installed, this is not a feasible procedure on the Internet. Thus, if a StochasticGeneratorServer is started in Internet mode it registers itself with a centralized Gatekeeper process, currently running on http://www.nmrshiftdb.org. If a client wants to perform a calculation it contacts the Gatekeeper to retrieve a set of idle machines on which it may perform the calculation (Figure 1). The Gatekeeper is thus only a mediator between client and server. The rest of the structure elucidation process involves only client and server.

The SENECA Client. Import of spectra and preprocessing for the server is the client's main task. A SenecaDataset, the data structure holding all the information associated with one particular structure elucidation problem, is represented by a single window in which the configuration parameters and spectroscopic information from different sources are organized as nested tabbed panes (see http://seneca.sourceforge.net for examples).

A typical setup for a CASE calculation starts with entering the molecular formula and some descriptive information for the dataset, followed by import of spectral data. The client can import peak picking lists from

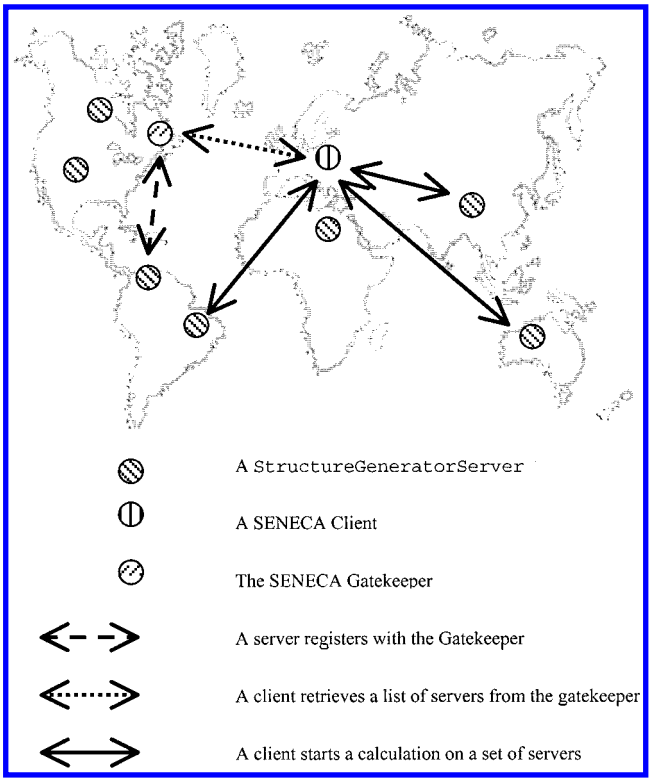

Figure 1. Distributed computing over the Internet with SENECA.

Bruker WinNMR and from the NMR processing software GIFA,[14] which is freely available for academic use (http://tome.cbs.univ-montp1.fr/GIFA). The third step is setting up the Judges, which provides an intermediate layer for the user to influence how the Judges interpret the underlying spectral data.

Finally, either a single StochasticGeneratorServer job is run on the client machine *or* the local subnet is scanned for available compute servers *or* a list of clients is retrieved from the gatekeeper for a calculation on the Internet.

I/O. Storage and retrieval of the data maintained by SENECA is performed in XML (see http://www.w3c.org). This would be an additional motivation for the spectroscopist to use SENECA, since it provides her with means to store and administer the spectroscopic information for a compound in a single place and in a highly portable, free, and human-readable data format.

The top level element ⟨SenecaDataset⟩ encapsulates structural information, for examples from result structures, encoded in CML,[15] NMR spectroscopic information coded in SpecML, an internally developed XML application which I propose as a basis for a standard way of representing spectroscopic information in XML, and some administrative information. Currently, SENECA has a clear focus on 1D and 2D NMR techniques. To be of general use it needs to be subjected to a wider discussion among spectroscopists. While a more extended discussion of this XML application is beyond the scope of this article, an self-explanatory example of a DEPT-135 spectrum coded in SpecML is available as Supporting Information, and a forum for discussion will be provided on the SENECA website.

It is noteworthy that due to the use of XML, any CML aware program (like JChemPaint,[16] JMol [http://www.openscience.org/jmol], or Jumbo [http://www.xml-cml.org]) can extract structural information from a SENECA dataset.

RESULTS AND DISCUSSION

The performance of SENECA's simulated annealing algorithm has been tested with a number of different annealing schedules and target compounds. The following section outlines a few typical structure elucidation runs, using as an example the sesquiterpene eurabidiol ($C_{15}H_{28}O_2$), isolated from the plant *Euryops arabicus*,[17] and the fungal metabolite monochaetin ($C_{18}H_{20}O_5$), a compound that served for this purpose in at least two earlier publications on CASE systems,[18] as well as polycarpol ($C_{30}H_{48}O_2$), a triterpene isolated from *Onychopetalum Amazonicum*.[19] The structures of these three compounds are shown in Figure 2.

The data for monochaetin were entered as listed by Christie and Munk.[18] They comprise 18 carbon chemical shift values and the corresponding DEPT data, 12 proton chemical shifts, 12 $^1J_{CH}$ correlations from CH COSY, 6 $^3J_{HH}$ correlations from HH COSY, and 28 ambiguous correlations from HMBC which can be interpreted as $^2J_{CH}$, $^3J_{CH}$, and less likely $^4J_{CH}$ or even farther reaching correlations. All these data were entered manually. Import of peak picking values, for example from Bruker WinNMR software, is possible for "real world problems".

The user starts the structure elucidation procedure by entering the molecular formula of the compound, $C_{18}H_{20}O_5$, and some optional administrative information, if desired.

The next step is to import all available spectroscopic data, as outlined above, into the 1D and 2D peak picking tables. Afterward, all carbon atoms in the molecule are labeled with carbon chemical shifts. If there are less carbon signals in the spectrum than carbon atoms in the molecule, shifts of signals with double, triple, or higher intensity are used multiple times, accordingly. During this automated process the number of hydrogen atoms bound to each carbon atom is assigned, based on inspection of the DEPT spectra. User

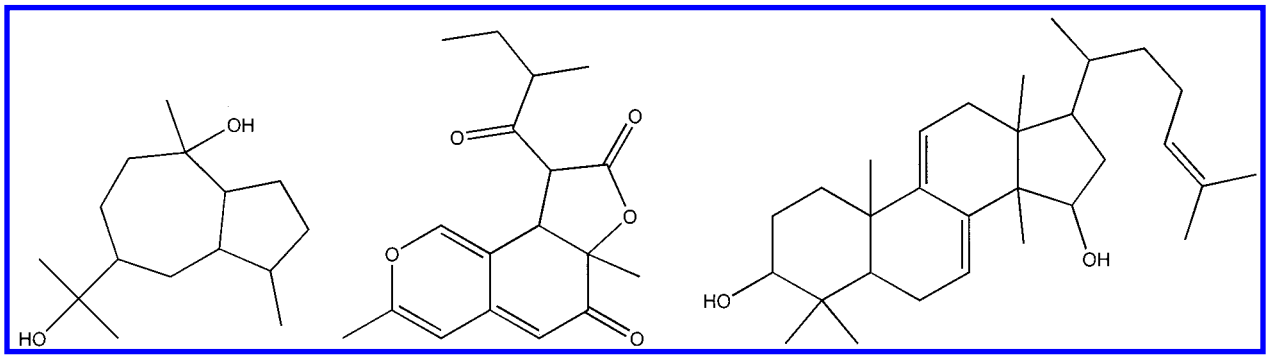

Figure 2. Structures of eurabidiol ($C_{15}H_{28}O_2$), monochaetin ($C_{18}H_{20}O_5$), and polycarpol ($C_{30}H_{48}O_2$).

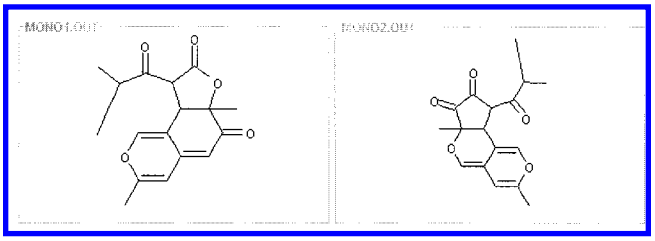

Figure 3. A LUCY calculation on monochaetin yields two result structures. Carbon atoms with chemical shifts $\delta_C > 100$ ppm were assigned as tricoordinate ACFs. Carbon atoms at 191 and 205 ppm were assigned to be carbonyl groups.

intervention is only required in case of ambiguity, for example if there are multiple possibilities for assigning hydrogens to heteroatoms.

The last action before starting the CASE run is to initialize the Judges. In the TwoDSpectrumJudges, like HHCOSY-Judge or HMBCJudge, for example, tables of relations between heavy atoms are produced from combined inspection of the $^1J_{CH}$ correlation data and the HH COSY for the first or the HMBC for the latter judge. Details of this process are described in the section "TwoDSpectrumJudge". This is also done by an automated procedure, which honors identical shift values for multiple carbon atoms as well as poorly resolved 2D spectra by assigning multiple relations between all pairs of atoms that might possibly be involved. Despite the automatism, I decided not to hide this procedure from the user because it is the most important step of abstraction in the whole process which should be transparent and fully configurable.

Once all desired judges are configured and activated, the user acquires a set of server nodes by asking a gatekeeper, doing a subnet scan or starting a local server on her machine, on which the CASE is then performed and monitored by the client. Upon completion of the calculation, the client retrieves the result structures from the servers performing an isomorphism filtering to eliminate duplicate structures and ranking them according to their scores. The structures can

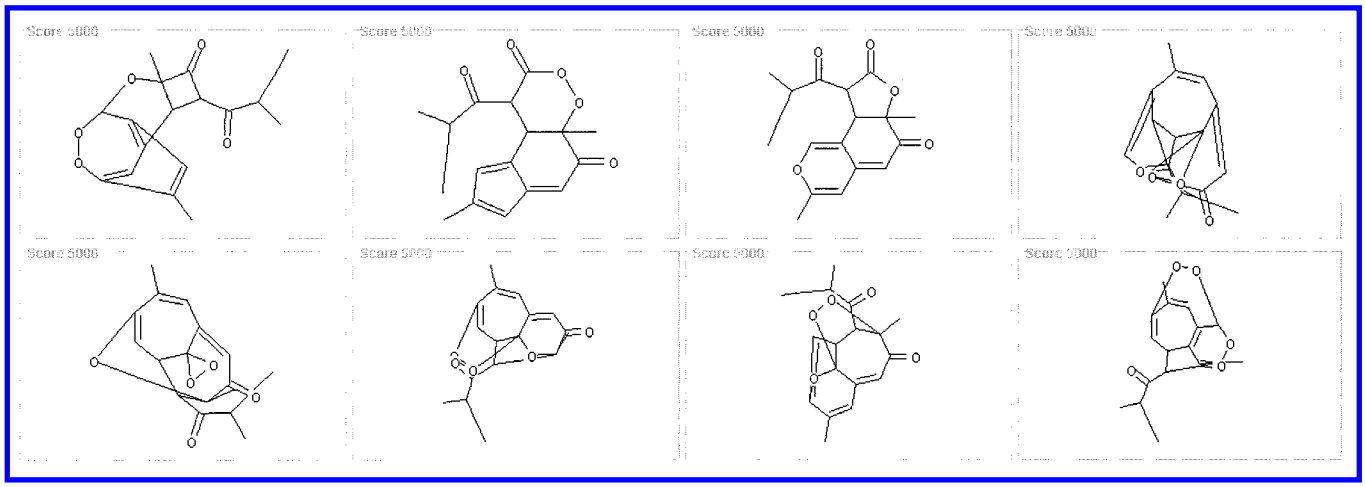

Figure 4. The eight structures with the highest possible score of 5000 points for the SENECA calculation on the monochaetin dataset using HHCOSYJudge, HBMCJudge, and HOSECodeJudge. If a FeatureJudge is used to introduce an additional penalty for bonds between heteroatoms, only the correct structure is found.

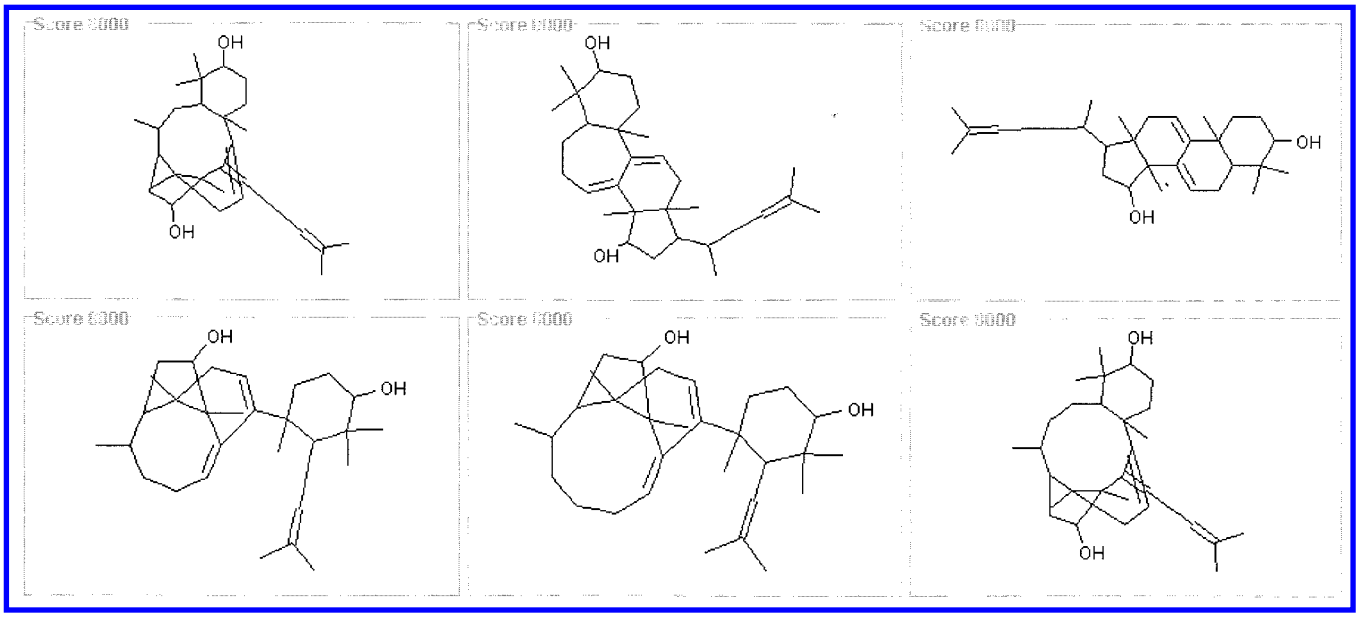

Figure 5. The six structures with the highest possible score of 8000 points for the SENECA calculation on the polycarpol dataset using HHCOSYJudge, HBMCJudge, and HOSECodeJudge. This is the full set of structures in agreement with the experimental data, as verified using the deterministic CASE program LUCY.

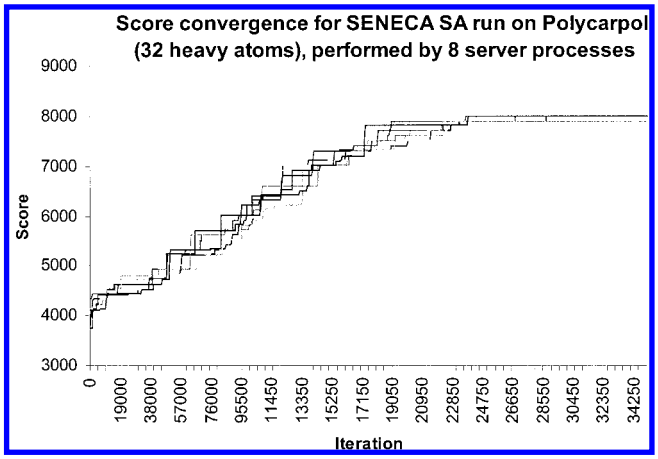

Figure 6. Convergence of the score sum (from HHCOSYJudge + HMBCJudge + HOSECodeJudge) in a CASE run on the triterpene polycarpol ($C_{30}H_{48}O_2$). The CASE run was performed in parallel on eight servers. The maximum possible score (see text) of 8000 was first encountered after about 220 000 iterations. Two servers did not reach the maximum score before the temperature convergence criterion was reached.

finally be inspected visually (see for example the screenshots in Figure 5) and stored in either CML[15] or MDL mol file format.[20]

Christie and Munk[18] report a varying number of result structures for their CASE run on Monochaetin, depending on whether carbon atoms in the chemical shift range between 105 and 116 ppm are treated as tricoordinate atom centered fragments (ACF) or whether both tri- and tetracoordination is accepted during the calculation. In the first case, they retrieve only one, the correct structure of monochaetin; in the latter case they find a set of six structures, of course containing the correct one.

LUCY[2] did not allow for mixed hybridization settings, and so the calculation was run using the less conservative settings with all tricoordinated ACFs for all carbon atoms with chemical shifts larger that 100 ppm. In this case, LUCY takes 19 s on a regular 600 MHz Personal Computer running Windows NT to find two structures in agreement with the input data, as shown in Figure 3.

A SENECA calculation with the same spectroscopic data was performed in parallel on eight server nodes (600 MHz commodity type PCs, running under Linux). Only three Judges, namely HHCOSYJudge, HMBCJudge and HOSECodeJudge, were used for scoring during the Simulated Annealing run. The calculations took about 9 min and converged on average after 250 000 iterations (generated structures). The eight structures with the highest possible score of 5000 for this data set is shown in Figure 4.

A data set for the triterpene polycarpol, which already served as an example in ref 2, comprises 30 carbon NMR signals, full DEPT information on CH_n multiplicities, 39 CH long-range correlations (HMBC), and 24 HH COSY correlations. Interestingly, the calculation converged after only slightly more time and iterations (12 min, 350 000 iterations), compared to the monochaetin data set. It yielded a set of six structures (Figure 5) with the highest possible score of 8000 points. As shown in ref 2, this is the full set of structures in accordance with the underlying spectral data. Figure 6 shows the convergence of the score for this calculation on all eight servers. Table 1 summarizes the data above including those of a third example, the sesquiterpene eurabidiol.[17]

Table 1: Running Time and Number of Iterations Needed for the SENECA CASE Runs of Three Selected Example Compounds of Growing Size[a]

name	molecular formula	calculation time		iterations
		LUCY	SENECA	
eurabidiol	$C_{15}H_{28}O_2$	29 s	5 min	90 000
monochaetin	$C_{18}H_{20}O_5$	16 s	9 min	250 000
polycarpol	$C_{30}H_{48}O_2$	33 min	12 min	350 000

[a] The calculation was performed using commodity type desktop PC's running at 600 MHz processor speed.

The above results show that the presented algorithm is able to search the full space of isomers, find the full set of correct solutions, but only needs to visit a fraction of the actual search space. It is difficult to make exact statements on the scaling because the exact number of isomers for the examples used is not known and because the natural differences in the spectral data describing the structure certainly influence the speed at which the correct solution is first encountered. However, the data in Table 1 are in agreement with a polynomial behavior, which, in contrast to the exponential growth of time for the LUCY calculation, strongly suggests that the SENECA system will be able to tackle problems too large for deterministic algorithms.

Two further advantages of the CASE scheme proposed in this article should be mentioned: First, besides its gentle scaling behavior it is the flexibility in accepting all kinds of contributions to the scoring function which makes the proposed scheme so compelling. Recently, for example, a fast method for measuring the natural-product-likeness of a compound has been proposed.[21] A judge working on this principle could, in the case of a natural product, on the fly narrow the result set in badly defined datasets with many results. A judge for drug-likeness would be another option.

Second, the scoring function itself makes the algorithm more flexible in honoring the exceptions to the rules of interpreting cross signals in 2D NMR spectra. In deterministic case systems, a cross signal in an HMBC spectrum, for example, is almost exclusively interpreted as either a two- or a three-bond, but not a four-bond, correlation between a carbon C-1 and a coupling proton H-2. In this case the ambiguity can be removed by constructing $N - 1$ unambiguous, alternative fragments (in a molecule with N heavy atoms) for each cross signal. These fragments comprise the directly bonded heavy atoms C-1 and X-2 (to which H-2 is bonded) and all combinations of C-1 and X-2 by inserting all $N - 2$ possible other heavy atoms as spacer. This has to be done for all HMBC cross-peaks. Four-bond correlations are rare but do occur, and in this case the algorithm above fails and the CASE system cannot find the correct structure. Generating all the possible combinations with two spacer atoms, on the other side, would significantly increase the number of solution structures as well as the calculation time and would not honor the rareness of four bond correlations in HMBC spectra.

In SENECA, this problem can be easily overcome by the scoring system. If an HMBC cross signal, for examples, can be explained by a two- or three-bond correlation in the result structure, 100 points are given, but only five points are granted in the case of a four-bond correlation. An HMBC signal from a $^4J_{CH}$ coupling will thus not lead to false-

negative results, but the structure will still be found and still be ranked highly within the set of other solution structures.

CONCLUSION AND OUTLOOK

It has been shown that CASE, based on 1D and 2D NMR data with only the knowledge of the molecular formula of the unknown compound, can be efficiently performed by stochastically searching constitution space guided by simulated annealing. Structures as large as triterpenes have been successfully elucidated. This is the standard size of the majority of problems that have been treated by deterministic methods in the literature. A great advantage of the new stochastic CASE system reported in this article is that very few assumptions are made prior to the structure elucidation run. Neither are larger fragments deduced from the spectral data in preprocessing run, nor are particular hybridization states assumed. Structures with an increasing percentage of heteroatoms impose as high a level of uncertainty in this case as they do in conventional structure generation based on NMR experiments. This is due to the lack of information that can be extracted from 2D NMR about the question whether a particular carbon atom is bonded to a heteroatom or not. Improving the precision of the HOSECodeJudge by employing larger spheres as well as the employment of other spectroscopic methods could be a way to tackle this problem. We are working on the assembly of a web-based open-submission/open-retrieval NMR shift database (www.nmrshiftdb.org) which, as soon as large enough, will serve as a basis for a free shift prediction tool.

The agenda for further development of SENECA includes the following: (1) finding better annealing schedules and evaluating the best weighting for the different score types during calculation, (2) implementing a scripting language for a more flexible control of the structure elucidation process, (3) improving the back-calculation of carbon NMR spectra in the HOSECodeJudge by the use of data assembled by NMRShiftDB, and (4) adding a deterministic structure generator as a reference system and for the exhaustive examination of small molecules. (5) It is an obvious move to compare the performance of the SA algorithm with other popular stochastic methods. We have implemented a structure generator based on a genetic algorithm and will report on this subject soon.

With the publication of this paper, SENECA will be made available on http://seneca.sourceforge.net as open-source software under an artistic license (see http://www.opensource.org for information) that will give researchers free access to the software as well as its source code and leave to me some semblance of control over where the software's development is heading. This will hopefully trigger a faster and more collaborative, and thus effective, development of the package as well as a broader and more extensive use by the community than has been the case with other CASE programs so far.

ACKNOWLEDGMENT

This work was funded by the Deutsche Forschungsgemeinschaft. My particular gratitude goes to Prof. Dr. W. Boland for his kind support and enduring encouragement.

Supporting Information Available: A SpecML encoding of the DEPT-135 carbon NMR spectrum of the α-pinene ($C_{10}H_{16}$) example data set in the SENECA distribution. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Munk, M. E. Computer-Based Structure Determination: Then and Now. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 997−1009. Nuzillard, J. M. Computer-assisted structure determination of organic- molecules. *J. Chim. Phys. Physico-Chim. Biologique* **1998**, *95*, 169−177. Jaspars, M. Computer assisted structure elucidation of natural products using two-dimensional NMR spectroscopy. *Natural Product Reports* **1999**, *16*, 241−247.

(2) Steinbeck, C. Lucy − A Program For Structure Elucidation From NMR Correlation Experiments. *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 1984−1986.

(3) Faulon, J.-L. Stochastic Generator of Chemical Structure. 2. Using Simulated Annealing To Search the Space of Constitutional Isomers. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 731−740.

(4) Metropolis, M.; Rosenbluth, A.; Rosenbluth, M.; Teller, A.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(5) Kirkpatrick, S.; Gerlatt, C. D. J.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.

(6) Guarnieri, F. Simulated Annealing. *Encyclopedia of Computational Chemistry*; John Wiley & Sons: Chichester, 1998; pp 2596−2599.

(7) Bremser, W. HOSE − A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.

(8) Bremser, W. Exspectation Ranges of 13-C NMR Chemical Shifts. *Magn. Reson. Chem.* **1985**, *23*, 271−275.

(9) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures − A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(10) Faulon, J. L. Isomorphism, automorphism partitioning, and canonical labeling can be solved in polynomial-time for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 432−444.

(11) Xu, Y.-j.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181−185.

(12) Figueras, J. Ring Perception Using Breadth-First Search. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 986−991.

(13) Shirts, M.; Pande, V. S. Screen Savers of the World Unite! *Science* **2000**, *290*, 1903−1904.

(14) Pons, J. L.; Malliavin, T. E.; Delsuc, M. A. Gifa V4: a complete package for NMR data-set processing. *J. Biomol. NMR* **1996**, *8*, 445−452.

(15) Murray-Rust, P.; Rzepa, H. S. Chemical markup, XML, and the Worldwide Web. 1. Basic principles. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 928−942.

(16) Krause, S.; Willighagen, E.; Steinbeck, C. JChemPaint − Using the Collaborative Forces of the Internet to Develop a Free Editor for 2D Chemical Structures. *Molecules* **2000**, *5*, 93−98.

(17) El-Sayed, A. M.; Al-Yahaya, M. A.; Shah, A. H.; Hartmann, R.; Breitmaier, E. Eurabidiol and other Sesquiterpenes from Euryops Arabicus. *Chemiker-Zeitung* **1990**, *114*, 159−160.

(18) Christie, B. D. The Role of Two-Dimensional Nuclear Magnetic Resonance Spectroscopy in Computer-Enhanced Structure Elucidation. *J. Am. Chem. Soc.* **1991**, *113*, 3750−3757.

(19) Hellwig, V. Einige Alkaloide aus Schizantus litoralis und ein Triterpen aus Onychopetalum amazonicum. *Diploma Thesis*, Bonn, 1994.

(20) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K.; Grier, D. L. et al. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(21) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245−1252.

CI000407N