Molecular Structure Elucidation Using Ant Colony Optimization: A Preliminary Study

Caroline Farrelly, Douglas B. Kell, and Joshua Knowles

Manchester Interdisciplinary Biocentre, The University of Manchester
Manchester, UK
j.knowles@manchester.ac.uk

Abstract. Identifying the structure of unknown molecules is an important activity in the pharmaceutical industry where it underpins the production of new drugs and the analysis of complex biological samples. We present here a new method for automatically identifying the structure of an unknown molecule from its nuclear magnetic resonance (NMR) spectrum. In the technique, an ant colony optimization algorithm is used to search iteratively the highly-constrained space of feasible molecular structures, evaluating each one by reference to NMR information on known molecules stored (in a raw form) in a database. Unlike existing structure elucidation systems, ours: does not need prior training or use spectrum prediction; does not rely on expert rules; and avoids enumeration of all possible candidate structures. We describe the important elements of the system here and include results on a *preliminary* test set of molecules. Whilst the results are currently too limited to allow parameter studies or comparison to other methods, they nevertheless indicate the system is working acceptably and shows considerable promise.

1 Introduction

Analytical chemists exploit a variety of spectroscopic techniques in order to gain an insight into the structure of unknown molecules. They use the molecule's exact mass, available from mass spectrometry, to reveal the empirical formula (e.g. C_4H_8BrF), and then study the molecule's spectral fingerprint to understand something about how these atoms are arranged. With NMR spectroscopy, patterns of chemical shifts can reveal information about local structures, from which it is (theoretically) possible, often after considerable toil, to infer the global molecular form.

Computer assistance for the task of *structure elucidation* has been available for decades now, initially as a means of helping to enumerate parts of the structural space so that chemists would be sure not to overlook any of the exponentially many possible forms. More recently, various AI techniques have been employed to automate the process further (see Section 5). For the most part, these techniques work by enumerating possible structures and then predicting the spectra of each one, which is then compared to the observed spectrum of the unknown molecule. This approach requires training machine learning methods to perform spectrum

M. Dorigo et al. (Eds.): ANTS 2008, LNCS 5217, pp. 120–131, 2008.

prediction, a science which is developing but still far from a solved problem. Moreover, the training process is intricate and time-consuming, and needs to be targeted to the particular kinds of molecules of interest. The quantitative comparison of observed and predicted spectra in these systems is also a nontrivial task which represents a further area under development.

In this paper, we investigate whether it may be feasible to tackle the structure elucidation problem *without* the use of spectral prediction methods. The approach we propose searches the space of possible structures iteratively using ant colony optimization [1], and evaluates candidate structures more directly by reference to a database of chemical shift patterns for known molecules. There is no explicit training necessary in our proposed method (in the sense of supervised learning), which potentially makes our system easier to update and less of a black box. From a machine learning perspective, the approach we use is similar to *lazy learning* [2]: we store our 'prior knowledge' in a fairly raw and uncompressed form and wait for a query before doing some work on the data to answer the query.

At the core of the system is a search of the candidate molecule space; the prior knowledge data is used mainly as an approximate evaluation function. Our motivations for choosing ant colony optimization as the search method are twofold. First, there are many constraints involved in building the structures and a constructive method such as ACO allows straightforward building of feasible solutions. Secondly, much of the structural information in a candidate structure relates to the order with which small modules (or substructures) are put together. Thus, we can treat the problem as a pseudo-ordering problem. We know that ACO is good at ordering problems from its successes in TSP, assignment, and scheduling applications [1]. In addition, some local enumeration of molecular structures is necessary to ensure all possibilities have been exhausted; and we know that combinations of ACO and local search tend to perform well (e.g., see [3,4]).

The rest of the paper is organized as follows. Section 2 formulates the problem that we tackle in this work, and relates it to other problems in machine learning and optimization. The problem is addressed by the approach we set out in detail in Section 3. Section 4 presents results from running the proposed method on a number of real NMR spectrum-to-structure problems. We discuss related literature on small molecule structure prediction from NMR in Section 5 and in Section 6 we summarise the initial findings presented here and look ahead to further developments.

2 The Spectrum to Structure Problem

The 'Spectrum to Structure Problem (SSP)' asks for the chemical structure of a molecule, given the molecule's spectral shift pattern and its empirical formula (EF). In the version of the problem we consider here, we are concerned with small organic molecules up to 500 molecular weight (MW) and the spectra are ^{13}C NMR spectra. The EF given denotes only the constituent atoms in the

molecule, not their arrangement. Different arrangements of the same constituent atoms are known as isomers; for even relatively small molecules there can be many isomeric forms, each giving rise to a slightly different NMR spectrum, e.g. the hydrocarbon C_9H_{16} has 1 902 isomeric forms. Furthermore, the number of isomers grows exponentially with the number of constituent atoms.

An example of two isomers and their spectra is given in Figure 1. Notice, we are concerned only with finding the 2D structure as represented by standard stick and ball diagrams. These structures have a one-to-one correspondence with the full chemical name of the molecule as given by the IUPAC nomenclature [5].

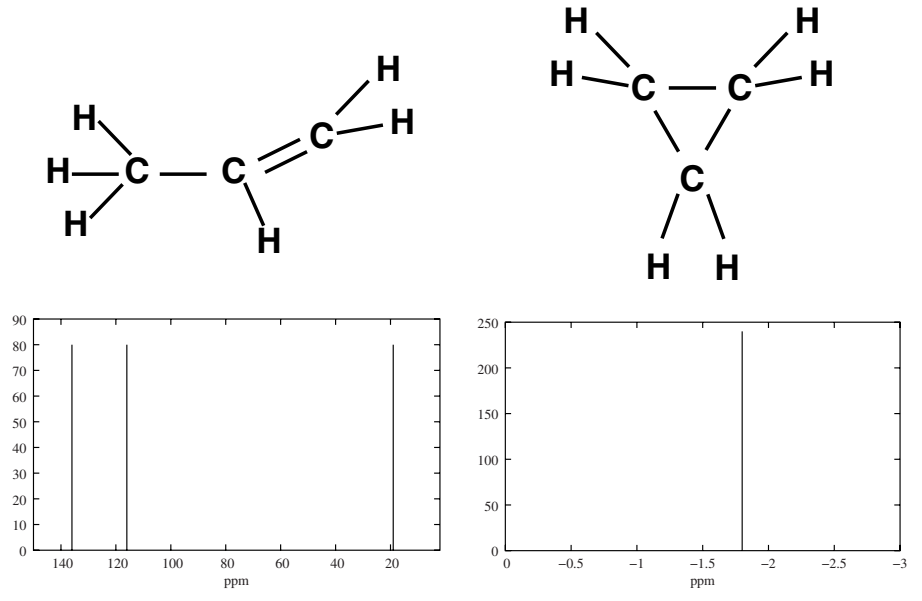

Fig. 1. Two-dimensional representations of the two constitutional isomers of the empirical formula C_3H_6 and their respective ^{13}C NMR spectra. In propane (left), three distinct shifts can be seen because each carbon atom's electronic environment is distinct. In cyclopropane (right), only one shift is visible because the 'view' from each carbon atom is identical.

In our version of the SSP, we assume that there is available some 'prior knowledge' in the form of a dataset giving the known 2D chemical structures and spectral shift values of organic molecules. Using this, we wish to infer the overall structure of an unknown molecule by identifying its likely substructural components — substructures that occur in molecules that exhibit similar NMR shifts to our unknown molecule.

A number of alternative formal definitions of the resulting inference problem could be given, based on measures such as 0-1 loss, or precision and recall (as used in classification problems). However, we choose here to allow that the inference method returns not one, but several attempts at inferring the structure. This

is because the problem is hard and a single attempted structure is unlikely to be correct, so measures of 0-1 loss would be unhelpful. Moreover, in practice, chemists would be happier to receive a small number of candidate molecular structures to investigate further, rather than one single answer that turns out to be inaccurate.

What we thus measure is the position of the true structure in a ranking (by internal fitness measure) of the candidate isomers generated by the inference method. Because our inference method is based on ACO, a stochastic metaheuristic, we must run the algorithm several times to evaluate performance, so to account for this, we report the overall rank of the true structure within a list (sorted by fitness) of all the unique structures generated by the ACO over the multiple runs performed. We also indicate the fraction of runs (out of those performed) on which the true structure is generated.

3 Ant Colony Optimization Approach

The method that we propose for tackling the SSP has three main steps.

1. **Data preparation:** Identify all *substructures* up to a given size that exist in the database; construct a matrix recording the frequency with which each substructure and spectral shift co-occur in the data. Split this matrix into two, one pertaining to smaller substructures and one pertaining to larger ones. (This whole step need only be done once for a given database of molecules. And if new data becomes available, the matrices can be *incrementally* updated by a trivial procedure.)
2. **Set constraints based on the query:** Once we have a query — an unknown molecule to identify — use its empirical formula to remove from consideration any substructures in the matrices that cannot be a part of the final structure, i.e., those that contain an atom that is not part of the given molecule and/or substructures that never produce any of the observed spectral shifts.
3. **Directed search:** Search for complete structures that match the empirical formula. Ants construct candidate structures from the smaller substructures identified in Step 1. The candidate structures are evaluated with reference to the larger-substructure frequency matrix from Step 1 via a maximum weighted assignment algorithm (see Section 3.6).

The underlying ant algorithm that we use for the directed search part is based on the MAX-MIN Ant System [6] (see Algorithm 1).

3.1 Data Preparation: Frequency Matrices

We are interested in building an approximate probability distribution over the substructures contained within the database. To obtain this information, the graph structure of each molecule can be decomposed into its subgraphs. Only subgraphs of limited size need to be found, where the size refers to the number

Algorithm 1. Ant colony optimization algorithm for structure elucidation

```
Input query: An empirical formula and its ^13C NMR spectral shift pattern
Prior knowledge: small substructures freq. matrix, large substructures freq. matrix
Constrain the search: Delete incompatible rows in the frequency matrices
Global best fitness ← 0
while Termination conditions not satisfied do
    for n = 1 to nants do
        Construct an ordered list of small substructures compatible with the empir-
            ical formula, using a pheromone matrix to guide choices
        Make all structures that are chemically possible from the ordered list
        for j = 1 to nstructures do
            Evaluate jth structure using a maximum weight assignment algorithm
        end for
        Record best fitness for this ant
    end for
    Record best fitness for this iteration
    Update global best fitness
    Update pheromone matrix with best fitness structure of the iteration
    if global best structure ≠ iteration best structure then
        Update pheromone matrix with global best structure
    end if
end while
Output: ranked list of candidate isomers and their estimated fitness values
```

of carbon atoms in the substructure. (All the non-carbon atoms bound to these carbons are also included). We use an algorithm that enumerates all valid substructures of sizes 2-carbon, 3-carbon and 4-carbon. Once this has been done for every molecule in the database, we are able to correlate substructures with spectral shifts. This is done by populating a matrix, which has rows representing substructures and columns representing shift frequencies (suitably binned into small value ranges) so that each element of the matrix records the number of co-occurrences of a substructure and a particular spectral shift. It is thus a representation of the joint probability of substructures and shifts (when correctly normalized).

We use this data in two ways in the ACO algorithm. We make one frequency matrix containing all 2- and 3-carbon substructures. These substructures are used as the solution components out of which the ants will construct full solutions. We make a second frequency matrix containing all the 4-carbon substructures only. This matrix is used to evaluate solutions (see Section 3.6).

3.2 Construction Graph Structure

The solution directly constructed by an ant is an ordered set of 2- and 3-carbon substructures, $s = \langle s_1, s_2, \ldots, s_k \rangle$, having a variable number of elements k. An ant begins with a partial solution $s^p = \emptyset$ and selects s_1 from the pool of available small substructures (with replacement) and adds it to s^p. The ant then makes the choice of s_2 adds it to s^p, and so on. The pool of available substructures is

updated after each choice to relect the constraint given by the empirical formula. The construction of a solution s is completed if the ant successfully completes a structure with the required empirical formula. It may also terminate construction, in the event that it is no longer possible to complete the empirical formula, which can occur if the addition of any substructure would result in exceeding the empirical formula in at least one atom type. In the case of terminating a solution without successfully completing it, the ant returns to the start of construction, setting $s^p = \emptyset$ and with the pool of available structures reset. An ant continues constructing solutions until it is successful.

The choice of substructure an ant makes at each step is mediated by both pheromone and heuristic information. Both of these sources of information help the ants to avoid making choices that lead to constructions ending in incomplete termination. A standard arc selection method is used [1], with the probability of selecting component c_{ij}, $i = 1, \ldots, k,\ j = 1 \ldots, |D_i|$ being given by

$$p(c_{ij}|s^p) = \frac{\tau_{ij}^{\alpha} \cdot \eta_{ij}^{\beta}}{\sum_{c_{il} \in N(s^p)} \tau_{il}^{\alpha} \cdot \eta_{il}^{\beta}}, \qquad \forall c_{ij} \in N(s^p)$$

where D_i is the domain of the decision variable (the set of substructures available to go in position i), τ_{ij} and η_{ij} represent the pheromone and heuristic information, respectively, and α and β are used to set the influence of these; $N(s^p)$ represents the feasible neighbours of the partial solution s^p.

3.3 Local Search: Translating Ant Solutions to Full Structures

The ordered list of substructures generated by an ant does not uniquely define an isomeric structure. This is because the substructures could be joined to each other in numerous ways. The ordering of the substructures is, however, intended to encode at least partially the preferred way in which the substructures should be joined. Thus, the solution encoded by an ant is interpreted as an instruction to join s_2 to s_1, then s_3 to s_2, and so on. But there are still numerous chemically valid ways in which this can be done that lead to different structural forms. These structural forms can be enumerated, given the ant solution. Therefore, an ant's construction is regarded as defining an ensemble of possible structures and the later evaluation of the ant solutions is done with respect to the best solution in the ensemble. Explicit details of the procedure for performing this enumeration of structures are given in [7]; space limitations prevent us from giving them here.

3.4 The Pheromone Matrix and Its Initialization

The pheromone matrix has m rows and k_{max} columns, where m is the number of substructures in the pool initially (after constraining), and k_{max} is the maximum possible number of substructures that could be needed to construct a valid isomer. It is simple to see that k_{max} is upper bounded by the number of carbon atoms in the empirical formula divided by two, since each substructure that we use to construct solutions has at least two C atoms.

An ant choosing substructural element s_j looks in the jth column of the pheromone matrix. The pheromone is thus on the nodes of the construction graph, and represents the relative desirability of selecting a particular substructure at a particular position in an ant solution (which as stated above represents an ordering of selected substructures). The pheromone matrix is initialized here with the maximum value τ_{max}, following Max-Min Ant System.

3.5 Heuristic Information

The heuristic information η_{ij} is given by

$$\eta_{ij} = \max(1, \mathrm{I}(c_{ij} \text{ completes EF}) \cdot 1000) \cdot \prod_{a \in A} h_{ij}^{a},$$

where A is the set of different atom types in the target empirical formula,

$$h_{ij}^{a} = \frac{\sum_{c_{il} \in N(s^p)} 1}{\sum_{c_{il} \in N(s^p)} \mathrm{I}(c_{il} \text{ contains atom type } a)}$$

and I(.) is the indicator function, which has value 1 if its argument is true, and zero otherwise. Thus, η_{ij} rewards a substructure c_{ij} if it contains an atom type a which is in the target empirical formula and if this atom is rare (or infrequent) in other available substructures. This encourages the picking of substructures containing rare atoms early on in solution construction, which helps prevent building candidate solutions that cannot be completed. The heuristic value of a substructure is further rewarded (by a factor of 1000) if its selection would complete the target empirical formula; this prevents making poor decisions towards the end of solution construction.

3.6 Evaluation Using the Maximum Weighted Assignment

To evaluate a candidate isomer, it is first mined for all its constituent 4-carbon substructures. A match between these larger substructures and those in the frequency matrix that co-occur frequently at similar spectral shifts would indicate a credible structure.

To assess the overall quality of these matches, we find the best assignment of shifts to substructures possible, and evaluate this assignment. Specifically, we have a set M of mined 4-C substructures and a set of observed shifts F. We have a weight matrix $W : M \times F \rightarrow R$ that stores the number of co-occurrences of each $m \in M$ and each shift $f \in F$ within the frequency matrix.

We would like to assign each shift precisely one carbon atom, but the substructures contain 4 carbons each. Therefore, we can allow each substructure to be matched with up to 4 shifts. To facilitate solving this as a standard bipartite graph matching problem, we can just copy each element of M four times to obtain an expanded set Q and expand our weight function to be $W : Q \times F \rightarrow R$, by simply repeating the weights four times. We now seek an assignment $g : F \rightarrow Q$ such that

$$\sum_{f \in F} W(f, g(f))$$

is maximized. This is a bipartite maximum weighted matching problem (or assignment problem) and can be solved by various methods including the Hungarian algorithm [8], though we used a restart hillclimbing method.

The solution to this problem here gives the most favourable interpretation of whether the set of substructures within the isomer could explain the shift pattern seen.

3.7 Pheromone Update

The elements in the pheromone matrix that appear in a solution to be rewarded (an iteration best ant or elite ant) are updated according to the following equation:

$$\tau_{i,j}(t+1) = (1-\rho) \cdot \tau_{i,j}(t) + \Delta\tau_{ij}^{best}$$

where $\Delta\tau_{ij}^{best} = 0$ if c_{ij} is not a component used in the best ant, and is otherwise the raw score derived from the weighted assignment problem described above.

Pheromones are forced to remain within the ranges set by τ_{min} and τ_{max}, by setting values below (respectively above) these to the respective bounding value.

4 Preliminary Experimental Results

Our experiments were conducted with the parameters of the ACO set as shown in Table 1(i). The basis of our experiments was a database of molecules compiled by us, as described in Table 1(ii).

The performance of the structure elucidation method is evaluated in two ways here. First, we examine if it can recover the structure of a molecule that is in the prior knowledge data itself. This is already a hard problem (and is NOT equivalent to testing on the training set in a classification/supervised learning task, because the space of *possible* structures that we search is still very large — much larger than our whole database of known structures, so we are not just learning class labels). These results are reported in Table 2.

Table 1. (i) Parameters of the ACO algorithm; (ii) Details of the database of known molecules

(i)

Parameter	*value*
nants	5
max iterations	80
τ_{min}	0.5
τ_{max}	10
ρ	0.01
α	1.0
β	1.0

(ii)

Training set info	
Number of molecules	2 873
Maximum MW	500
Minimum MW	50
Total number of atom types	16
Number of 2C and 3C substructures mined	2 881
Number of 4C substructures mined	5 926

Secondly, we verify the performance on molecules not in the initial knowledge-base. This is achieved here by 'holding out' certain molecules we wish to test from contributing to the frequency matrices. Due to some limitations of our data-sets, we can only do this for two molecules at present (see Table 2, bottom).

Table 2. Test results on a range of small organic molecules. The target molecule is found in all cases and in almost all runs. Often the approximate fitness of the true structure means that it is ranked highly amongst the other candidates. Bottom: results on hold-out data.

No. of carbon atoms	Molecule name (IUPAC convention)	Empirical formula	No. of isomers as enumerated by [9]	Rank by fitness (total no. of unique isomers gnrtd.)	Number of runs target gnrtd. / total runs
4	1-bromo-2-fluorobutane	C_4H_8BrF	12	1st (2)	27/27
	4-aminobutanenitrile	$C_4H_8N_2$	633	1st (27)	17/17
	1-methoxypropan-2-ol	$C_4H_{10}O_2$	28	1st (12)	18/18
5	(1E)-1,2-diiodopent-1-ene	$C_5H_8I_2$	88	3rd (8)	14/14
6	1-(allyloxy)propan-2-ol	$C_6H_{12}O_2$	1313	17th (396)	10/11
	1-propoxypropan-2-ol	$C_6H_{14}0_2$	179	6th (127)	8/8
	1,1'-dithiodipropane	$C_6H_{12}S_2$	timeout	1st (66)	15/15
7	1-butoxypropan-2-ol	$C_7H_{16}O_2$	463	15th (292)	20/20
Hold-out data results:					
6	1-propoxypropan-2-ol	$C_6H_{14}0_2$	179	2nd= (127)	5/5
7	1-butoxypropan-2-ol	$C_7H_{16}O_2$	463	10th (292)	4/4

The results reported in Table 2 are currently limited by a couple of factors that have prevented a larger study. These are that: (i) our system of joining substructures cannot currently generate ring structures, which means that a significant fraction of structures cannot be tested yet; and (ii) at several points, our code calls proprietary software to convert between different representations of chemical structures (namely, SMILES strings and MOL files), which creates a substantial computational bottleneck that prevents us from testing the larger structures in our database. We are working to overcome both of these factors, which are certainly not inherent problems of the system.

Despite the limitations, the results are positive on the cases we have tested, with true structures being correctly recovered in all cases, and often ranked highly by the assignment method compared with other structures generated. On the hold-out data, the ACO system worked at least equally well when these isomers were removed from the prior knowledge database as when they were in it. Much more testing is required to understand the effect of the distribution of isomers stored in the database on performance; but this initial test indicates that it is not necessary to have seen the molecule before to predict its structure using our system.

5 Related Work on Structure Elucidation

In comparison to the number of applications available for spectrum prediction, the field of structure elucidation is relatively small and immature. Most attempts to address this issue are built upon an expert system with an inherent rule base.

A common feature is the requirement for an empirical formula. From this, all possible isomers are generated and a spectrum is predicted for each one, allowing for similarity ranking against the original query spectrum. Although this significantly narrows the search space, typically thousands of isomeric forms may remain. If there are significant distinctions between the spectra, the structure corresponding to the top ranking spectrum can be taken as the structure causing the experimental spectra. However, if several top-most ranking spectra are very similar, further analysis may be required. It should be noted at this point that the both the accuracy of spectrum prediction and similarity ranking are of primary importance in structure elucidation, because the larger the margin of error in these, the more likely it will be that the predicted structure will be incorrect.

There are several factors to be taken into account during ranking, including matching the number of nuclei visible in the spectrum, chemical shifts values and scalar couplings. It can be difficult to determine corresponding shifts between predicted and experimental spectra, especially where multiple shifts occur within a small separation. A study has highlighted how matrices can be used to detect optimal matches between experimental and predicted spectra [10]. The first two expert systems developed for this area, CONGEN [11] and GENOA [12], generated isomeric forms, from which a specialist would select a likely structure. Both systems required considerable human interaction in forming lists of favoured or unlikely fragments, but GENOA allowed fragment overlap within the isomers constructed.

A more modern trend in structure elucidation applications is to utilize several different spectral types in order to perform elucidation, for example multiple dimensions, element types or analytical techniques. This rapidly reduces the chemical search space and facilitates ranking. Programs such as CHEMICS[13], X-PERT [14], and StrucEluc [15] use such supplementary data to determine specific libraries and rules which should be accessed in order to improve search results.

A more unusual approach is taken by the program Genius [16], which uses a genetic algorithm for structure generation. A neural network is used to categorise the electronic environment of each carbon in an isomer and then to predict its spectrum. Using a GA for structure generation means not all isomers need be initially generated, potentially narrowing the search space. The level of similarity to the query spectrum determines which chromosomes are allowed to mate and reproduce into the next generation. Runs can be stopped either by correct structure determination (matching chemical shifts), time limits (after a set number of generations), or accuracy limits (when chemical shifts lower than those in the experimental spectrum are achieved).

6 Summary and Future Work

A system for tackling the spectrum-to-structure problem based on ACO has been presented. The system does not use expert rules, nor does it rely on predicting spectra from structures; instead, iterative heuristic search is combined with the use of a knowledge-base of identified structures and their characteristic spectra. On the data used to test the system here, it produced sets of proposed structures that contained the true structure in all cases, even when the space of possible isomers was large (i.e. containing over a thousand feasible structures). On several occasions, the true structure was the isomer ranked highest by the system. Moreover, the set of structures that a chemist may regard as likely candidates can potentially be reduced further, by taking account of the pheromone trail information, rather than considering every structure generated. Testing has obviously been very limited to date so it is not possible to draw any more than preliminary conclusions from this. However, we are encouraged by these results to continue further investigations.

The system now needs to be extended to tackle different molecular forms, such as rings, which it is currently incapable of identifying (see [7] for more details). We need to test the system further and compare it with alternative approaches, including existing spectrum-to-structure methods and simple baseline approaches. Such testing will require us to gather more high-quality NMR spectral data for similar and larger molecules, to allow much larger studies to be done, with more quantitative reporting of success rates as well as computation times.

The spectrum-to-structure problem will continue to be an important one in the pharmaceutical and systems biology arena. There is a growing need for fast identification of molecules that have been manufactured artificially, such as candidates for active pharmaceuticals (drugs), or naturally-occurring molecules that have never been characterized before, such as many of the metabolic products of biological cells [17,18]. The work started here may eventually allow us to build systems that are more scalable — requiring less human input and expertise and less time-consuming training — than currently available ones.

Acknowledgments. Many thanks to Dr Lee Griffiths (Astra Zeneca, Alderley Park) and Dr Bryn Roberts (formerly Astra Zeneca, Alderley Park) for advice on NMR spectroscopy and for providing access to chemical shift data. Caroline Farrelly was supported by a CASE studentship from Astra Zeneca and EPSRC, UK. Joshua Knowles is supported by a David Phillips Research Fellowship from BBSRC, UK.

References

1. Dorigo, M., Stützle, T.: Ant Colony Optimization. MIT Press, Cambridge (2004)
2. Aha, D.: Lazy Learning. Kluwer Academic Publishers, Norwell (1997)
3. Stützle, T., Hoos, H.: MAX–MIN Ant system and local search for combinatorial optimization problems. Meta-Heuristics: Advances and Trends in Local Search Paradigms for Optimization, 313–329 (1999)

4. Gambardella, L., Dorigo, M.: An Ant Colony System Hybridized with a New Local Search for the Sequential Ordering Problem. INFORMS Journal on Computing 12(3), 237–255 (2000)
5. http://www.chem.qmul.ac.uk/iupac/
6. Stützle, T., Hoos, H.: MAX-MIN Ant System. Future Generation Computer Systems 16(8), 889–914 (2000)
7. Farrelly, C.: From Spectrum to Structure Using Machine Learning. PhD thesis, School of Chemistry, University of Manchester, UK (2008)
8. Munkres, J.: Algorithms for the Assignment and Transportation Problems. Journal of the Society of Industrial and Applied Mathematics 5(1), 32–38 (1957)
9. MOLGEN Tool, http://molgen.de/?src=documents/molgenonline
10. Griffiths, L., Bright, J.: Towards the automatic analysis of ^{1}H NMR spectra: Part 3. Confirmation of postulated chemical structure. Magn. Reson. Chem. 40, 623–634 (2002)
11. Carhart, R., Smith, D., Brown, H., Djerassi, C.: Applications of artificial intelligence for chemical inference. XVII. Approach to computer-assisted elucidation of molecular structure. Journal of the American Chemical Society 97(20), 5755–5762 (1975)
12. Carhart, R., Smith, D., Gray, N., Nourse, J., Djerassi, C.: GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures. J. Org. Chem. 46, 1708–1718 (1981)
13. Sasaki, S., Kudo, Y.: Structure elucidation system using structural information from multisources: CHEMICS. Journal of Chemical Information and Computer Sciences 25(3), 252–257 (1985)
14. Elyashberg, M., Martirosian, E., Karasev, Y., Thiele, H., Somberg, H.: X-PERT: a user-friendly expert system for molecular structure elucidation by spectral methods. Analytica Chimica Acta 337(3), 265–286 (1997)
15. Elyashberg, M., Blinov, K., Williams, A., Martirosian, E., Martin, G.: Application of a New Expert System for the Structure Elucidation of Natural Products from the 1D and 2D NMR Data. J. Nat. Prod. 65(5), 693–703 (2002)
16. Meiler, J., Will, M.: Genius: A genetic algorithm for automated structure elucidation from C-13 NMR spectra. Journal of the American Chemical Society 124(9), 1868–1870 (2002)
17. Kell, D.: Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discovery Today 11(23-24), 1085–1092 (2006)
18. Kell, D.: Metabolomic biomarkers: search, discovery and validation. Exp Rev Mol Diagn 7(4), 329–333 (2007)