Similarity and the Beilstein Information System: Searching for Concepts with Current Facts

Martin G. Hicks

Beilstein Institute, Varrentrappstrasse 40-42, 6000 Frankfurt 90, FRG

Received June 23, 1992

The theme of similarity in organic chemistry is one of increasing importance. There are at least two reasons for this: first, the chemist should be enabled to ask questions of an expert system in a way that is closely related to how he thinks, people think in images and the analysis of which, in terms of similarity, is fundamental to our thought processes; and secondly, he should be provided with better models and therefore a better understanding of chemistry. Two interconnected similarity measures are used in the Beilstein Information System: the Beilstein System itself, which defines the ordering of compounds within the Beilstein Handbook and is therefore a classification system, and the Lawson Number, which is derived from a computer algorithm used to classify compounds according to the Beilstein System and is a means of similarity searching on Beilstein Online and Beilstein Current Facts in Chemistry CD-ROM. Chemical concepts are defined here in terms of an overlap in the similarity hyperspace between structural similarity and property similarity. The ability to search for chemical concepts using Current Facts is demonstrated.

INTRODUCTION

Artificial Intelligence versus Artificial Stupidity. Expert systems should be able to give intelligent answers to intelligent questions, but the definition of intelligence varies depending on whether you are human or a computer. Computers are precise, do routine work without becoming exhausted, and sort and handle large amounts of information; things that are often described as nonintelligent tasks. Humans operate in a different way; their processing or thinking, while potentially analytical, is often based on images and the comparison of one image with another that is similar.

It is important for a chemist to be able to get an answer from a chemical information expert system to the following question: "I want to see all molecules that are similar to this one". Similarity is context dependent and will accordingly have a variety of definitions—whether these are relevant for chemical, physical, or physiological properties or molecular structure or shape or not, the computer is expected be able to provide intelligent answers to our questions.

Many current systems have, as an antecedent, one or another classification system. Classification is used to order large sets of information so that humans, and nowadays computers, can handle them. The aim is to be able to find something again. Classifying compounds in groups allows a system to offer the user a certain degree of similarity searching; the disadvantage is that neighboring groups in the classification space often have nothing in common, and the fundamental directive "give me *all* compounds similar to this one" cannot easily be followed. Examples of classification systems are the Beilstein System[1] and, to some extent, many substructure search systems,[2] which use, for example, fragment codes to describe and, hence, classify molecules. However, these inflexible and, from the point of view of the user perhaps unintelligent, classification-based systems can also be used in ways not originally intended and enable the chemist to retrieve some or most of the compounds that are regarded as similar.

Similarity and Chemical Concepts. Similarity in chemistry[3] can generally be regarded in two ways; either from the macroscopic viewpoint in which similarity is defined in terms of chemical, physical, or physiological properties based on ensembles of molecules or from the microscopic viewpoint where similarity is defined in terms of molecular or atomic descriptors. Concepts can be defined as resulting from the overlap of these two views to provide a definition of macroscopic properties in terms of microscopic descriptors (Figure 1).

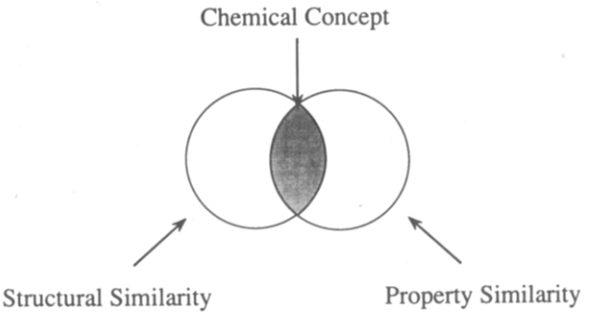

Figure 1. Definition of a chemical concept.

BEILSTEIN SIMILARITY MEASURES

Organic compounds are ordered in *Beilstein's Handbook of Organic Chemistry* according to the Beilstein System, which is a measure of chemical structure similarity based on structure segments. Thus compounds are sorted in the Handbook not according to name, molecular formula, or chronological order but according to a measure of molecular similarity. Browsing on either side of a particular reference in the Handbook gives access to similar compounds. The Beilstein System has been implemented on a computer, and the algorithm produces as additional output the Lawson Number.[4] A Lawson Number is associated with each structure in the Online and Current Facts databases and is, therefore, a readily available search feature.

The Beilstein Information System includes in its structure-oriented databases chemical, physical, and physiological information. The amount, type, and frequency of this information will be described in the context of the Current Facts CD-ROM database.[5] Since this contains information from the new literature, this is the information density toward which the database will tend with time. Current Facts also has the advantage of being a large database that is locally stored and is therefore readily available for the trial and error type of searching often necessary to extract concepts from a database.

0095-2338/92/1632-0631$03.00/0

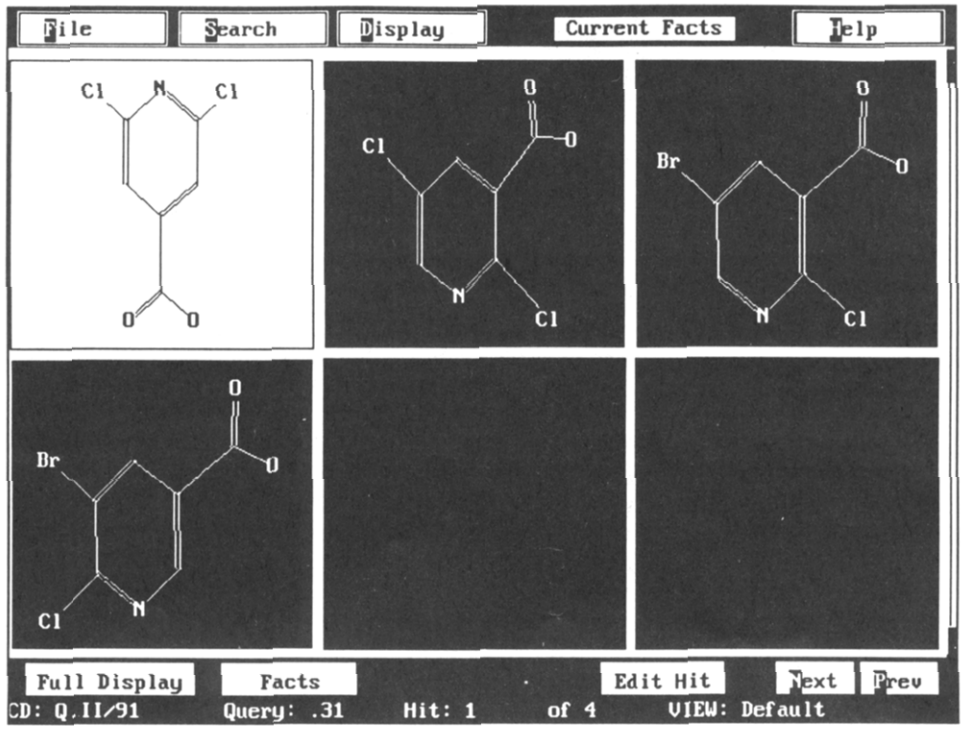

Figure 2. Dihalopyridinecarboxylic acids.

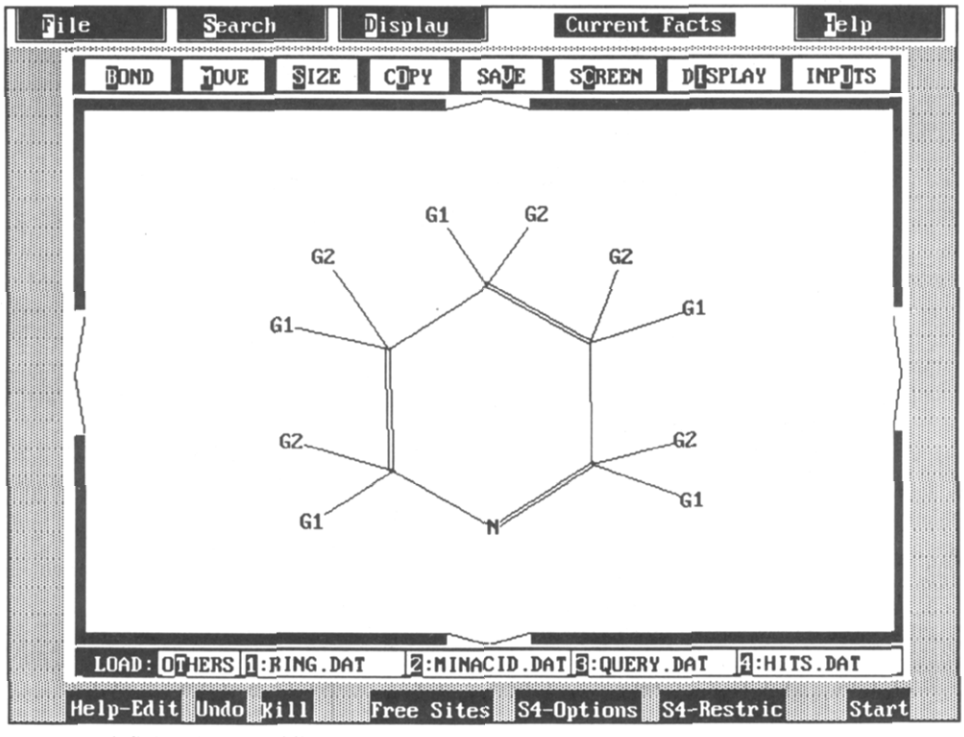

Figure 3. Markush structure defining dihalopyridinecarboxylic acids.

BEILSTEIN SYSTEM

The Beilstein System is a classification system for ordering organic compounds in *Beilstein's Handbook of Organic Chemistry* so that "similar" compounds are found close together. The system has been described fully elsewhere.[1] Each compound is given a system number, between 1 and 4720, which is assigned according to several rules which are briefly described:

Registry Compounds. These are the index compounds and are divided into three main groups: acyclic, isocyclic, and heterocyclic. A further subdivision is then made according to whether any of the following functional groups are present: hydroxy, oxo, carboxylic acids, sulfinic acids, sulfonic acids, selenic/selonic/telluric acids, amine, hydroxylamines/hydrazines, azo compounds, diazonium compounds, etc. The multiplicity and type of the functional groups, together with the degree of saturation and number of carbon atoms, define the exact position in the system.

Functional Derivatives. These compounds are placed after the equivalent registry compound in the index.

Formal Hydrolysis of Functional Derivatives. Functional group derivatives are formally hydrolyzed into their structure segments; for example, esters are hydrolyzed to acids and alcohols. The segments are then given their system numbers,

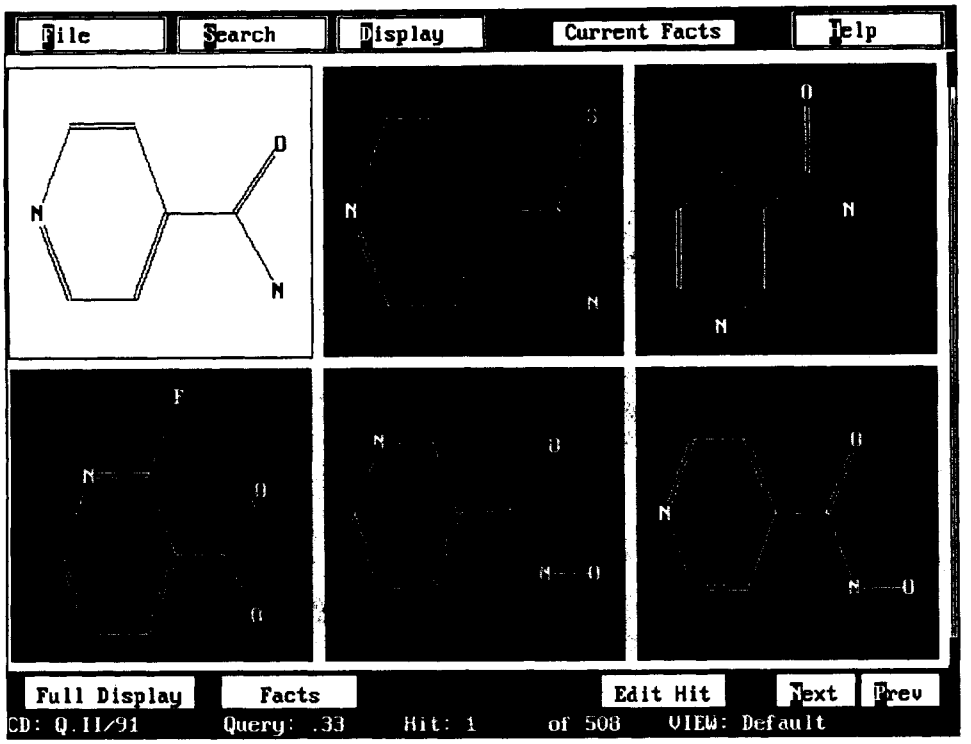

Figure 4. Results of a Lawson Number range search.

and the compound is indexed according to the segment with the highest system position.

Substitution Products. The following are defined as substitution products: F, Cl, Br, I, NO, NO_2, N_3, and N_5. Thus, compounds containing one or more of the above groups are indexed following their unsubstituted parents.

Chalcogen Analogs. Organic sulfur, selenium, and tellurium compounds are defined as being analogs of the corresponding oxygen compounds and are indexed after the oxygen compound.

Thus, the Beilstein Handbook has a very well defined classification system. It has the advantage that like compounds are found together and that the same compound will be found in the same place in all of the editions of the handbook.

LAWSON NUMBER

The Institute has been changing its operating procedures over the last few years and has been switching over to computerized production methods. These computer methods have been designed to relieve, wherever possible, the scientific personnel of the repetitive, time-consuming tasks. The obvious starting point was in the cataloging of the compounds for the Handbook. Computers are ideally suited to processing and sorting large quantities of data; thus, the first goal was to develop an algorithm that could classify compounds according to the Beilstein System.

This program has been in successful service in the Institute for some years. A spin-off from this program is the Lawson Number,[4] which is essentially a hash code based on the Beilstein System. It is a number between 0 and 32 767 and is, therefore, roughly a factor of 8 larger than the Beilstein System Number. The Lawson Number should, however, be regarded as being principally a classification code that can be "misused" as a similarity search tool.

Groups of compounds classified as being similar by the Beilstein System have similar Lawson Numbers. Taking a small range around a particular number generates a hit list of similar compounds, just as if one had browsed in the Handbook. The Lawson Number is present in the Online and Current Facts databases. Unlike the Beilstein System Number where only the highest number is used, all of the Lawson numbers derived from all the hydrolysis segments are stored in the Lawson Number field (LN) in the databases.

Table I. Subset of Data Field Occurrences in Current Facts Database

field code	field name	occurrences
BRN	Beilstein Registry Number	283 627
RN	CAS Registry Number	68 669
SO	Beilstein source	56 348
LN	Lawson Number	283 591
CN	chemical name	156 088
AN	autonom name	123 177
PRE	preparation	196 531
INP	isolation from natural product	4982
MP	melting point	77 160
BP	boiling point	9143
DEN	density	1140
RI	refractive index	3147
ORP	optical rotatory power	21 839
NMRA	NMR absorption	125 333
IRA	infrared absorption	71 840
EAM	electronic absorption maximum	16 493
MS	mass spectrum	28 993
IP	ionization potential	343
VP	vapor pressure	96
SLB	solubility	276
BF	biological function	21 851

SUBSTRUCTURE SEARCHING AND THE LAWSON NUMBER

Depending on the breadth of the search, substructure searches and Lawson Number searches can give the same results; for example, consider the search for all dihalopyridinecarboxylic acids. The Lawson Number search for these is as follows:

LN = 26334 AND ELC = 6

For precise control over the Lawson Number search results, the query is usually best combined with other structural characteristics. In this case, unless a further restriction is

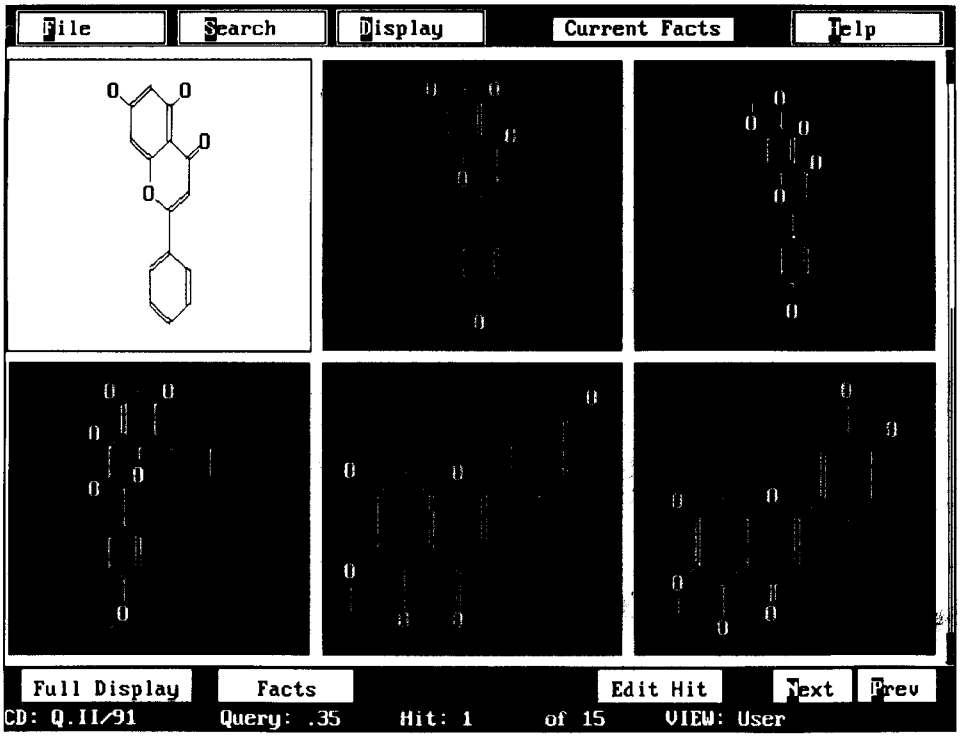

Figure 5. Results of the search BF = influenza* AND INP.

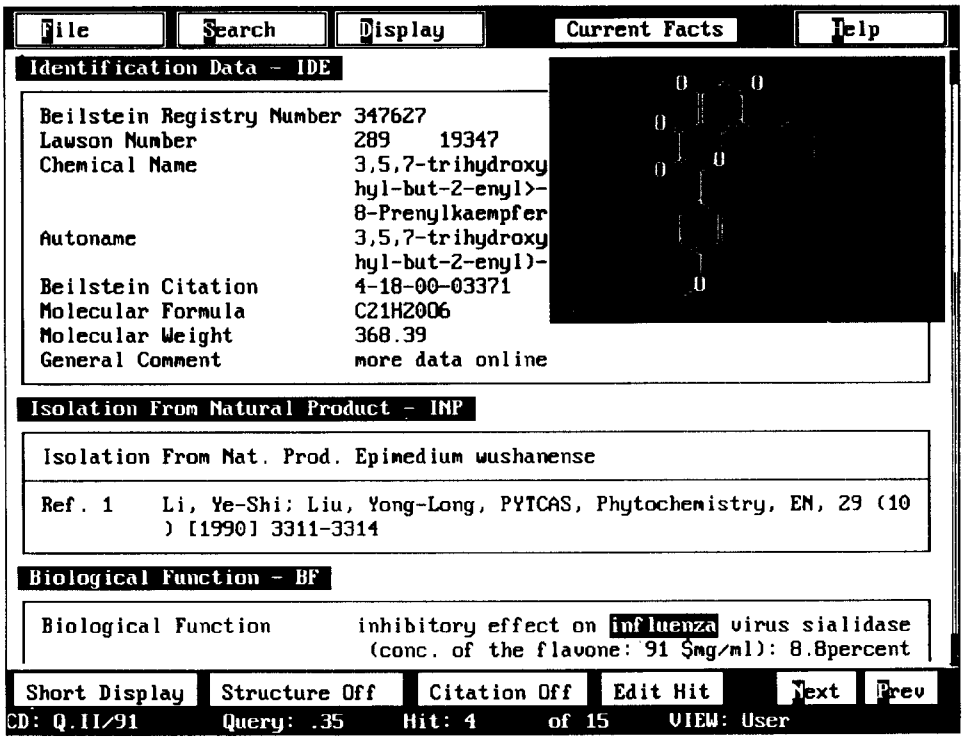

Figure 6. Full display of a compound from Figure 5.

applied, all esters would also be found. Applying the restriction Element Count (ELC) = 6 ensures that they are excluded. The results are shown in Figure 2. In order to retrieve the required hits with other data sets, further restrictions concerning N or Cl, for example, may also have to be applied.

The same answer set can be retrieved by searching for the Markush structure in Figure 3. Here G1 is defined as any halogen occurring with a frequency of 2 in the molecule, and G2 is a carboxylate group, with a frequency of 1.

This was a fairly narrow search, when the search boundaries are extended the advantages of a similarity measure become apparent. Searching for the Lawson Number range LN = 26328–26335 retrieves a hit list of 508 compounds that are, within the bounds of the LN, similar to each other. The first six are shown in Figure 4. To carry out the same expansion using a substructure search is very difficult; allowing maximum free sites on all atoms tends to loosen control over the degree of similarity. The only satisfactory way would be to use a list of generic groups; however, this often takes a prohibitively long time to input. Thus, the use of a similarity measure, in this case LN, has clear advantages when the search boundaries are relaxed.

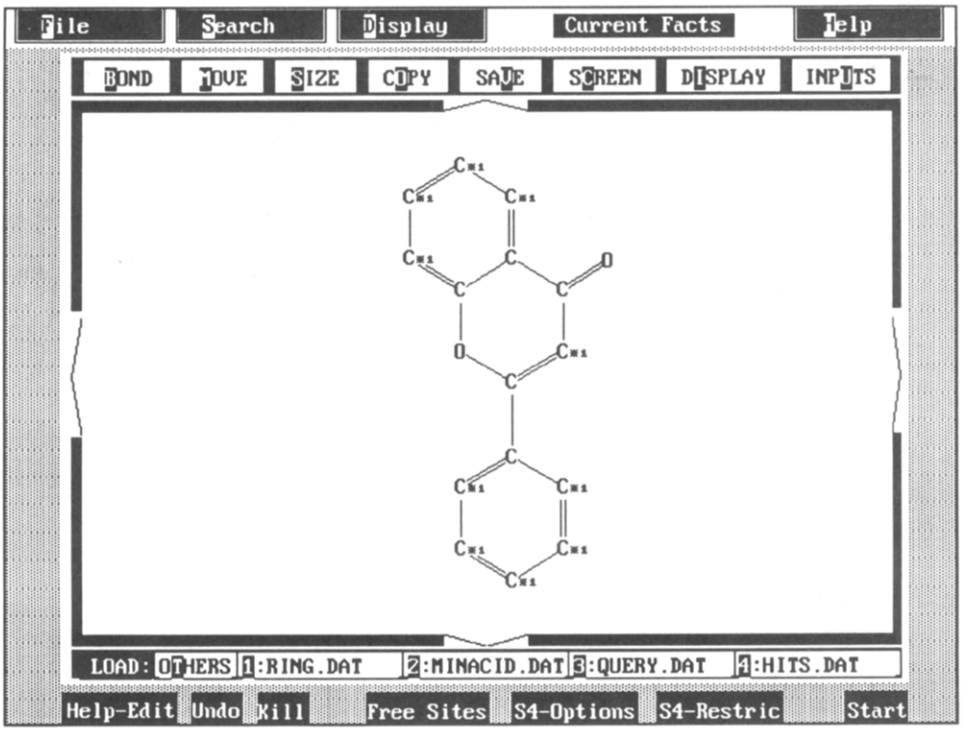

Figure 7. Generic flavone substructure.

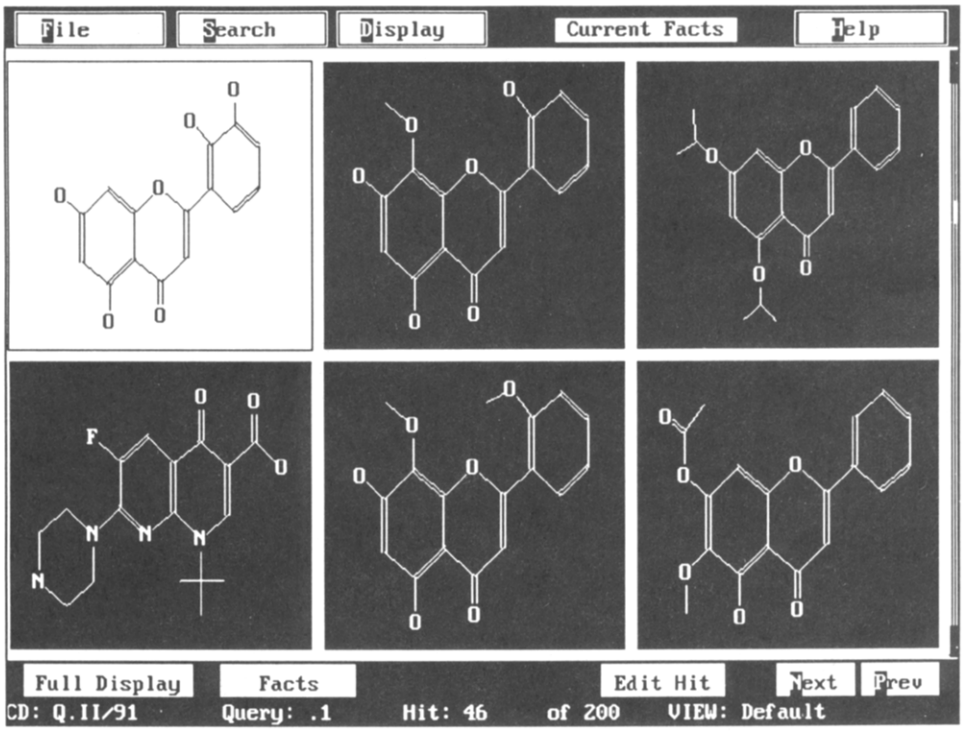

Figure 8. Results of a search for BF = influenza*.

COMBINING DATA AND STRUCTURE SEARCHING

If the previous example from the Lawson Number search is used in combination with a search for boiling points, a hit list of 10 structures with measured boiling points is retrieved. Depending on the exact use intended, this could be just sufficient to describe a series of compounds in terms of their physical data. A statistical analysis of the Current Facts database, which needs to be multiplied by about 15–20 to extrapolate to the Online database, provides an overview of the practicality of using factual information to define sets of similar compounds. A subset of the fields is shown in Table I.

Nearly 80% of the compounds are new, that is, they have not been mentioned in the Beilstein Handbook nor do they have a previously assigned CAS Registry Number. This can be seen from the occurrences of the SO and RN fields. That over 60% of compounds have at least one preparation is an indicator of the importance of this information. Systems that search for reactions, preferably by means of similarity measures, are of corresponding importance.

While some physical data fields, such as MP and BP, are frequently cited and, as we have seen, can often produce enough information for further use in, for example, calculation programs, other fields that are of perhaps more importance

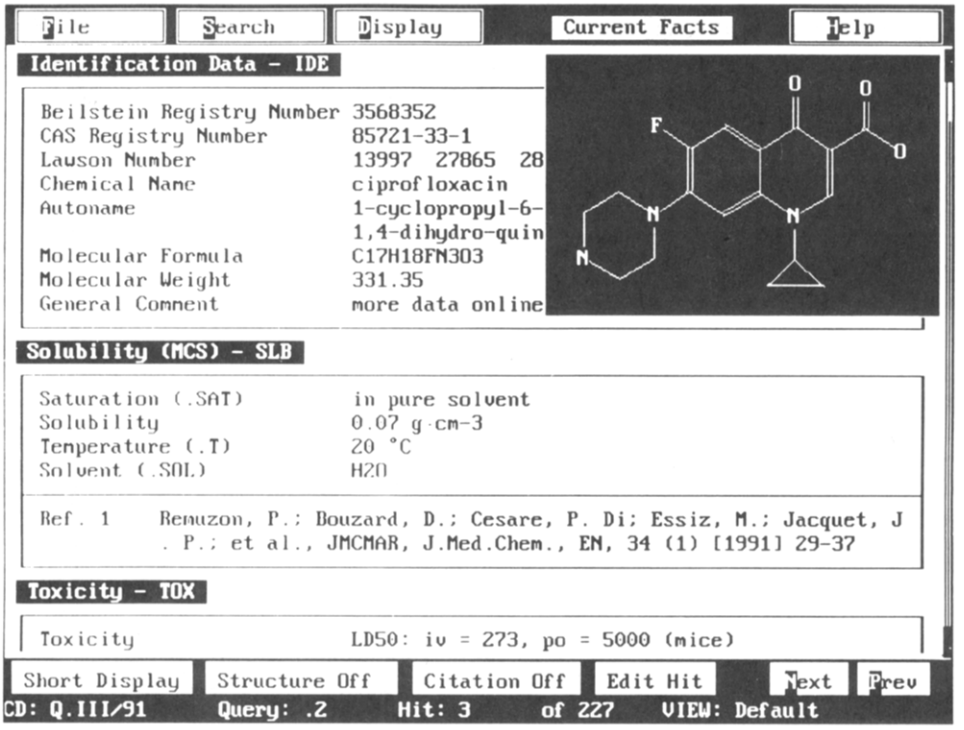

Figure 9. Ciprofloxacin.

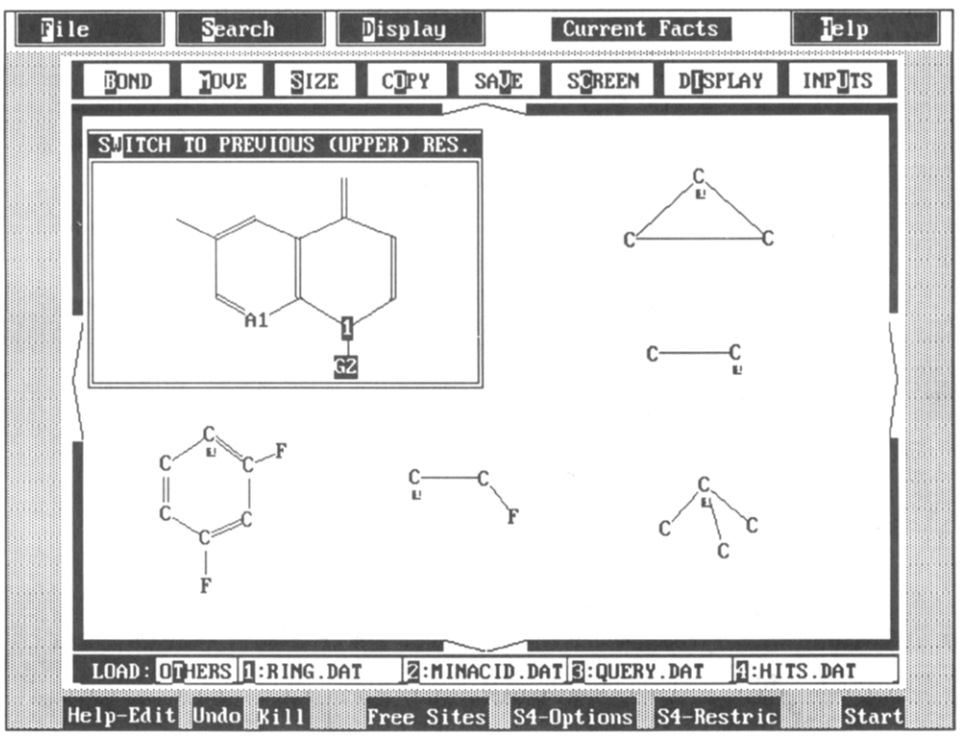

Figure 10. Generic fluoroquinolone.

for such programs are too sparsely filled in the Current Facts database to be used for defining sets of similar compounds. It is often the case that, to get enough information to describe one or more series of compounds thoroughly, the online database should be used. The best strategy is to first carry out the search on the CD, where the are no search charges, and then, after checking and perhaps modifying it, online.

CONCEPT SEARCHING

A chemical concept defined previously as being formed in the similarity space from the overlap of structure and physicochemical properties is something which is often elusive, and since it is not explicitly present in the databases, it cannot be searched for directly. Thus, the concept is formed by intellectual analysis of the subset produced by the intersection of two or more hit lists. There is no guarantee that the result is any more than a pointer to the right direction; a starting point for further investigation.

Current Facts is a large database containing current factual information which runs on a PC. This means that it is possible to carry out investigative searches which would be too expensive online. The following example demonstrates the use of the Current Facts database to find a substructure which confers biological activity on a complete structure.

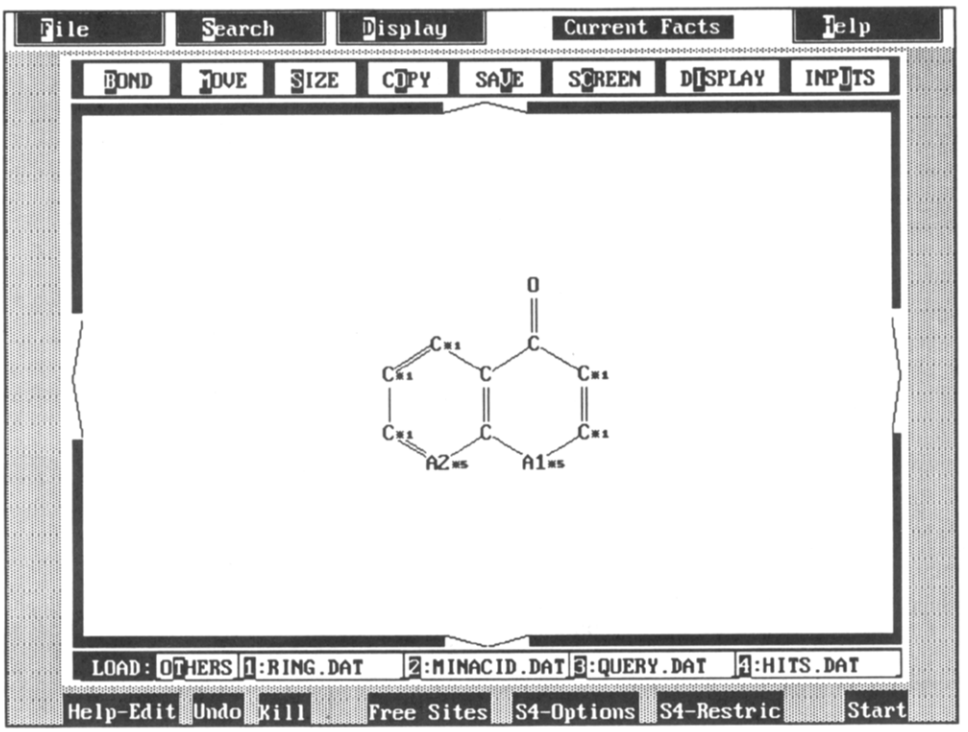

Figure 11. Generic substructure describing flavones and fluoroquinolones.

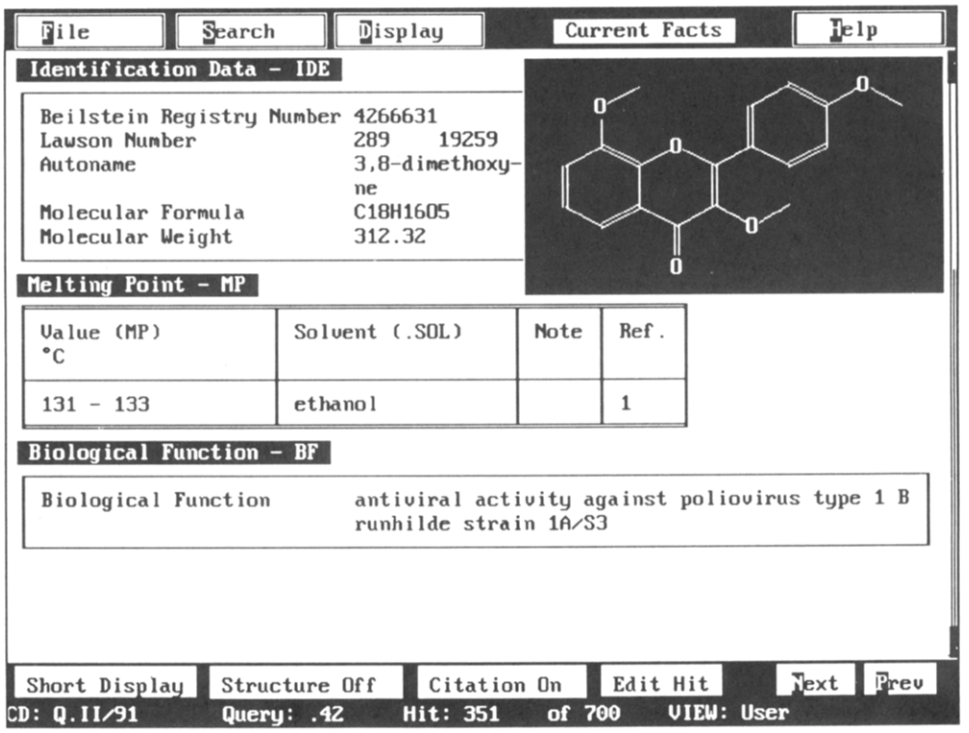

Figure 12. Example of an active compound.

The starting point was to find out if, in the previous year, any compounds were discovered which were both active against an influenza virus and were isolated from natural products. The following search was carried out:

BF = influenza* AND INP

BF is the field code for biological function, and INP is the field code for isolation from natural products. The asterisk in the search term "influenza" forces right truncation, thus enabling both influenza and influenzae to be retrieved; the lack of a search term after INP means that a "field availability" search is to be carried out, thus the presence of this field will be searched for.[5]

The patients of today frequently express a preference for "natural medicines". This applies to over-the-counter and prescription drugs. Market pressures have therefore driven pharmaceutical companies to find active natural products. Fortunately, nature provides a great treasure chest of compounds, the isolation and testing of which is significantly less expensive than their synthesis.

The search resulted in 15 hits. Examination of the hit list, part of which is shown in Figure 5, shows a great degree of similarity in the compounds found. Since they all come from one citation, this is only to be expected. The inherent information, that flavones (2-phenylchromones) are potentially

active, is exactly the type of starting point that we have been looking for. One of the hits is shown in full in Figure 6.

A generic flavone structure is easily derived (Figure 7). When searched for, the resulting hit list of 683 compounds is then intersected with the search term BF = influenza* and gives a hit list of 79 structures, containing compounds with measured activities but which are not necessarily isolated from natural products. Since a Current Facts CD-ROM contains only the information from one year's publications, it could well be that the isolation was previously published. If this is an important criterion, then a search in the online database is also necessary.

In a matter of minutes, a list of compounds has been compiled which meets the starting criteria. This could be end of the search, but it is interesting to see what happens when the search boundaries are extended as follows:

BF = influenza*

A hit list of 200 compounds is retrieved. Browsing through the list, the set of compounds shown in Figure 8 was found. In addition to the previously retrieved flavones, a fluoronaphthyridone was, among others, also present. This can be seen as a modified fluoroquinolone antibiotic, of which ciprofloxacin (Figure 9) is the best-known example. Taking this as a starting point for a new lead, a generic ciprofloxacin fluoroquinolone was defined (Figure 10). This is a fairly narrow search profile allowing only cyclopropyl, ethyl, phenyl, and isopropyl substitution at the 1 position. A total of 431 hits was retrieved, showing the importance of this compound in contemporary medicinal chemistry.

Thus, two pathways have been examined, and it is possible using a generic structure shown in Figure 11 to unite them. When position 1 is defined as a generic atom to be taken from the list of either O or N and position 8 as either N or C, we have a structure that describes both fluoroquinolones and flavones. Intersecting this list of structures with a list derived from a BF "field availability" search, a list of 351 compounds is retrieved. The example in Figure 12 shows that one of these compounds is active against the poliomyelitis virus. This generic structure defining the structural similarity together with the knowledge of the potential activity against viruses and bacteria defines the concept. This is obviously a simple chemical concept that requires further refinement to give a useful model. It is, however, a good starting point and has demonstrated how easily one can be defined, given powerful search tools and a large enough database.

CONCLUSION

Large databases make it necessary for the chemist to have a wide variety of tools at hand to enable him to retrieve the required information. Given intellectual support from the chemist, the existing search tools provide excellent access to large databases such as that of Current Facts. However the intellectual work necessary to arrive at the required information becomes harder with increasing file size, thus tools that support the intellect, such as similarity or concept searching tools, would provide major steps forward in this area.

ACKNOWLEDGMENT

I thank Dr. Douglas Maass for helpful discussions and for his invaluable assistance in the translation of this paper.

REFERENCES AND NOTES

(1) *How To Use Beilstein*; Springer-Verlag: Berlin, Heildelberg.

(2) For descriptions of substructure search systems see: (a) Hicks, M. G.; Jochum, C. Substructure Search Systems. 1. Performance Comparison of the MACCS, DARC, HTSS, CAS Registry MVSSS, and S4 Substructure Search Systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 191–199. (b) *Chemical Structures*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988 and references cited therein.

(3) For further reading on similarity in chemistry see: (a) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990. (b) Willett, P. *Similarity and Clustering in Chemical Information Systems*. Research Studies Press: Letchworth, Hertfordshire, England, 1987. (c) Relevant papers in this issue and references cited therein.

(4) Lawson, A. J. Chemical Structure Browsing. In *Chemical Structure Information Systems: Interfaces, Communication and Standards*. Warr, W. A., Ed.; American Chemical Society Symposium Series No. 400; American Chemical Society: Washington, DC, 1989; pp 41–49.

(5) Hicks, M. G. Beilstein Current Facts in Chemistry: A large chemical database on CD-ROM. *Anal. Chim. Acta* **1992**, *265* (2), 291–300.