SOFTWARE PACKAGE FOR THE INPUT AND STORAGE OF STRUCTURAL DATA ON THE IBM PC

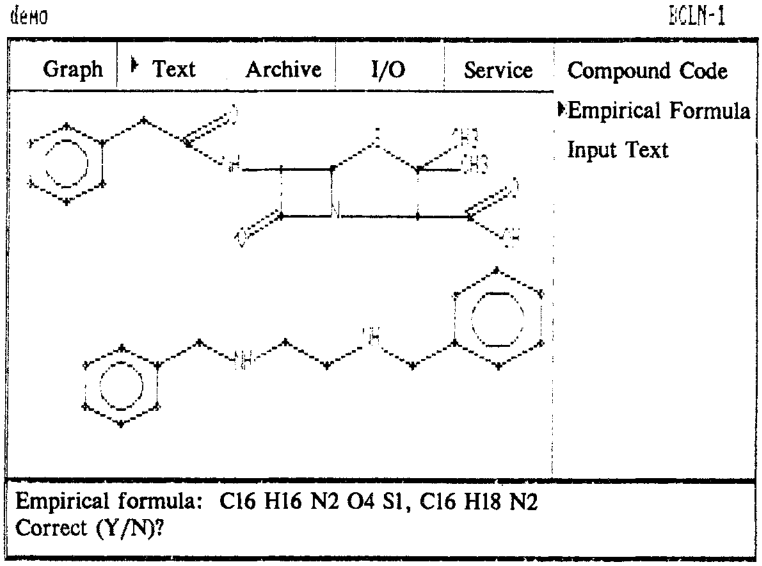
A. A. Yanik

UDC 541.6

One of the main problems in chemical informatics is the creation of effective management software for data from chemical investigations on the basis of modern methods of data storage and processing. The solution of this problem calls for developing techniques for representing the data on a compound in a computer and the creation of a technology for constructing chemical databases that are suitable for solving problems in both information science and applied science. The specific details of investigations in this area are determined by the existence of a special class of information-containing objects, i.e., structural formulas of chemical compounds. The highly effective processing of the influx of data in chemical information systems would be impossible without the use of machine methods for the input of chemical structural information.

One of the possible versions of such a system is the Computer Registry of Organic Compounds (CROS). The approaches used to create this system and its various versions were discussed by us several years ago in [1]. The CROS system is a specialized highly efficient graphics processor for the input, storage, and output of data on chemical structures and text. The system was written in the form of a software package in Microsoft FORTRAN 4.0 and operates in personal computers such as the IBM PC/XT or PC/AT with a CGA or EGA graphics adapter and a mouse. The system provides for:

1. The interactive input of graphic structural data and pertinent text (the name, synonyms of the name, selected properties, etc.);
2. Monitoring of the correctness of the input of the structural information with the use of dictionaries of permissible labels for the atoms and functional groups, which are included in the system;
3. Additional possibilities for creating a graphic image in the form most convenient for use;
4. Storage of the graphics data and text on external magnetic media with the aid of the archive system of CROS;
5. Import/export of data in the archive system for the exchange of information with other archive systems (in the framework of CROS) and for solving practical problems (for example, the calculation of topological indices).

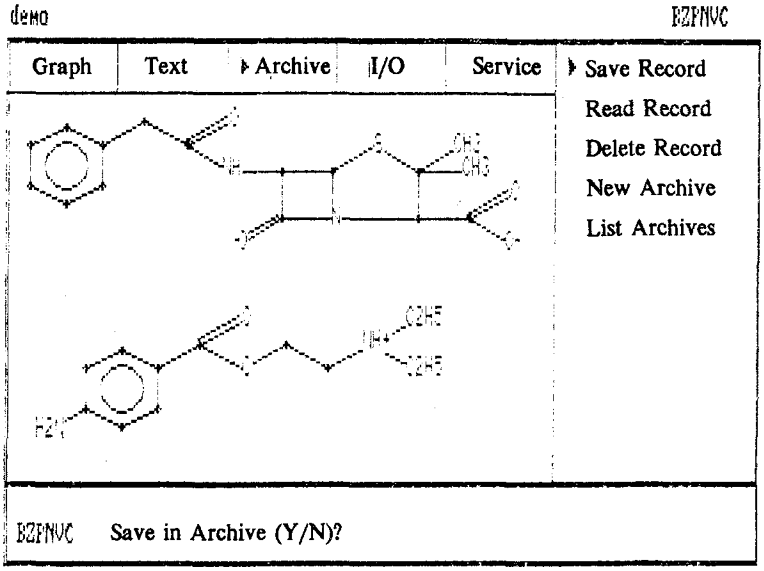


```
Empirical formula:  C16 H16 N2 O4 S1, C16 H18 N2
Correct (Y/N)?
```

Fig. 1.

"Medstatistika" Scientific-Research Institute, Ministry of Health of the USSR. Translated from Zhurnal Strukturnoi Khimii, Vol. 32, No. 2, pp. 157-159, March-April, 1991. Original article submitted July 24, 1989.

0022-4766/91/3202-0287$12.50 ©1991 Plenum Publishing Corporation



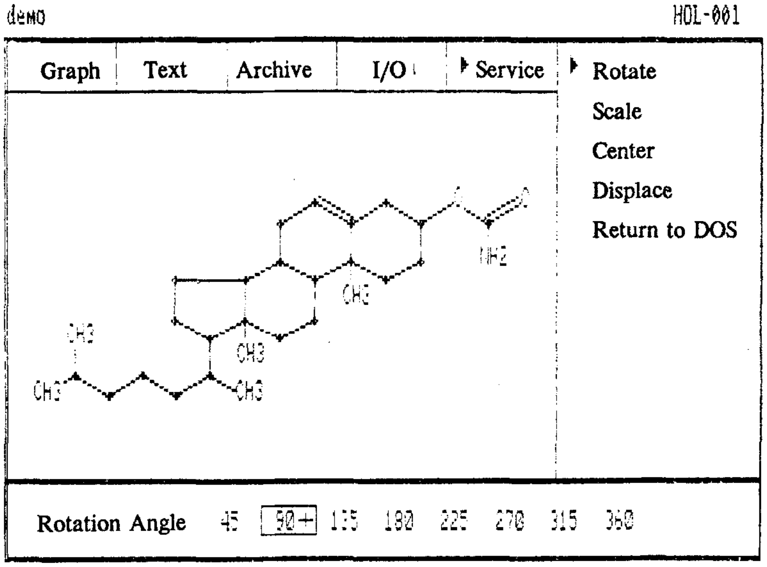
BZPNVC Save in Archive (Y/N)?

Fig. 2.

A multiwindow graphics user interface, which operates with the use of "prompts," was developed for convenience in controlling the system. Figures 1-3 present schematic representations of the display screen at the time of the performance of various procedures.

During the development of the CROS system, special attention was focused on the problem of formalizing chemical data and finding the most efficient methods for the input of chemical structural data from the point of view of computer graphics.

A system of conventions (standards) for representing structural formulas has been devised on the basis of the selection and unification of data which characterize chemical structures. On the one hand, these conventions ensure the representation of structural formulas in traditional forms used by chemists, and, on the other hand, they make it possible to utilize the machinery of graph theory for processing the internal representations of the chemical-structural data. Two standards were adopted and used to develop CROS: a basic standard and an expanded standard. The basic standard requires explicit indication of all the nonhydrogen atoms in the structural formula, and the expanded standard allows the use of an extensive set of functional groups and other means to simplify input. When the program for converting chemical-structural data from the expanded standard to the basic standard is present, the system of conventions adopted permits the use of different systems



Rotation Angle 45 90 135 180 225 270 315 360

Fig. 3.

for the input of chemical-structural data, which may be based not only on machine graphics, but also on various alphanumeric notations, such as the Wiswesser Line Notation, for creating chemical databases.

During the creation of CROS, special investigations associated with the development of the optimal (with respect to productivity and the minimization of errors) methods for organizing the input of chemical-structural data were carried out, and a number of non-Soviet systems, for example ChemBase, ChemFile, ChemSmart, and HTSS [2], as well as the experience of the scientific teams in Novosibirsk, Moscow and Riga [3-6], were studied. A scheme for the input of chemical-structural data by "assembling" the image of a structural formula from primary objects was adopted. "Rings of assigned size" and "atoms" were selected to serve as the latter. These primary objects are then "pasted together" with the aid of other primary objects in the form of "bonds of assigned types" and a procedure for finding the nearest neighbors. As a result of the development of special software, it was possible to achieve equal expenditures of labor (number of necessary operations) for the input of the primary objects. An isolated ring, a spiro ring system, and a condensed ring system are constructed on the screen in one operation, which is confined to pressing the key on the mouse.

We propose using the CROS system to create large data files in the area of the chemistry of organic compounds.

LITERATURE CITED

1. V. I. Goisa, V. E. Ivanov, A. P. Suchkov, and A. A. Yanik, Khim.-farm. Zh., No. 1, 97-102 (1987).
2. D. E. Meyer, ACS Symposium Series No. 341, American Chemical Society, Washington, D.C. (1987), Chap. 4, pp. 29-36.
3. V. B. Muchnik, R. S. Nigmatullin, and A. L. Osipov, NTI, Ser. 2, No. 8, 6-11 (1985).
4. N. S. Zefirov and S. S. Trach, "Use of computers in chemical investigations and molecular spectroscopy," in: Abstracts of Reports to the 7th All-Union Conference [in Russian], Riga (1986), pp. 17-19.
5. A. M. Kachalkov, S. G. Molodtsov, and V. I. Smirnov, Ibid., pp. 32-33.
6. A. B. Rozenblit and V. E. Golender, Combinatorial-Logical Methods in the Design of Drugs [in Russian], Zinatne, Riga (1983).