ANTIBIOTIC IDENTIFICATION: A COMPUTERIZED DATA BASE SYSTEM

Sir:

Antibiotics play important roles in the chemotherapy of infectious diseases and of cancer. Therefore it is the primary concern of many laboratories to discover other new antibiotics with different or improved properties. In discovering novel antibiotics it is essential that already known antibiotics are recognized early in the research process. Early recognition will ensure that time is not wasted in the pursuit of leads which result in the isolation of an already known antibiotic. The approach for early identification of known antibiotics varies from laboratory to laboratory. Some laboratories use paper chromatographic methods or thin-layer gel classification systems to compare partially purified, potentially new antibiotics with known reference compounds[1,2]. Another approach is to compare the unknown antibiotics' characteristics with the characteristics of known antibiotics which have been coded on edge-punched cards. A system of this type was described by J. BERDY[3,4]. Other systems, like computer assisted identification of substances comparing one or more characteristics of the unknown (mass spectra, IR spectra, *etc*) with the data on known substances.

The Chemotherapy Fermentation Laboratory of the Frederick Cancer Research Center, which was assigned the task of isolating new antibiotics with potential anti-tumor effects, required a system capable of identifying known compounds based on the chemical and physical characteristics of partially purified crude fractions. A computer system was designed to assist in the identification by operating on a comprehensive collection of characteristics of reported antibiotic compounds.[5] The computer system contains unique features which make it particularly suitable for this and similar applications. Unlike most data base systems, no limits are placed on the number and type of entries which can be entered for a compound. Because the system requires only that each set of values for an antibiotic compound be preceded by an identifier that defines those values, the user can create new identifiers and add their values. The data base is coded in a self-defining manner, so that it is natural for the research personnel involved to use their own terminology. The system allows for missing values in value sets and for tolerance ranges on numeric values. A search routine is incorporated into the system that allows for the matching of the characteristics of the unknown compound against those of the compounds in the data base; selecting those compounds in the data base most nearly matching the unknown. The system also produces a concordance of the value in the data base along with the compound identifiers of the antibiotics with the same value.

To establish the data base, characteristics of the antibiotics in the above-mentioned BERDY punch card system were recorded in a format acceptable to the computer system. The BERDY data base contains data obtained from the literature on more than 5,000 antimicrobial and antitumor agents. A subset of this data contains the compound characteristics that are particularly important for the early identification of the compound and were selected to be included in the computerized data base. Several other characteristics, such as the formula, the antibiotic classification number and the molecular weight, were included to make the listings of the data base useful as a reference tool. The characteristics which are included in the computerized data base are shown in Table 1.

The data base may be used to identify unknown compounds in one of two ways: (1) by running a search program that compares the unknown's characteristics to the data base, selecting those antibiotics with characteristics most nearly matching the unknown, or (2) by manually matching the unknown's characteristics against the concordance of the values and characteristics in the data base.

The search program accepts the characteristics of an unknown compound which are coded similar to the data base entries and compares them to the data base. Exact matches between the input search characteristics and the compounds in the data base are not required or expected. The values in the search string for an identifier are compared one for one with the values in the data base for the same identifier. A value in the search characteristics is considered to match a data base value if it is contained in the data base

Research sponsored by the National Cancer Institute under Contract NO1-CO-25423 with Litton Bionetics, Inc.

Table 1. Characteristics included in the antibiotic data base

Antibiotic name, synonyms	Chromatography
Chemical type*	Stability
Chemical formula	Test organisms
Elementary analysis	Toxicity data
Antibiotic code number**	Antitumor/antiviral effects****
Producing organisms	Isolation methods:
Molecular/equivalent weight	filtration
Appearance-physical characteristics***	extraction
Optical rotation	ion-exchange
Ultraviolet spectra	absorption
Solubility	chromatography
Qualitative chemical reactions	crystallization

* and ** According to the BERDY classification system[3].

*** Color, crystal structure, *etc.*

**** *In vitro* and *in vivo* characteristics.

or if the tolerance ranges on numeric values intersect. The program contains an algorithm to weight the search characteristics according to the order in which they were entered and to compute a score for each data base compound which reflects the degree of similarity between it and the search characteristics. After searching the data base, the compounds most nearly matching the unknown compound are displayed. The complete listing of the selected compounds may be used to confirm the identity of the unknown, or confirm that the compound does not exist in the data base and may be a previously undiscovered unreported one.

Confidence in the identification process may be justified only if the data base covers a large percentage of the reported compounds. The BERDY data base is the largest known collection of antibiotic data covering at this time over 5,000 compounds. The computerized subset of the data base covers the same compounds and contains over 1.5 million characters of data with 111,000 individual values or value sets. The data base system is periodically updated. A reference file has also been compiled which contains information such as the compound structure and IR spectra which are not on the computerized data base.

At the time of this writing, the characteristics of products from seven crude fractions with significant antitumor activity have been matched against the BERDY data base, using the computer program described above. The program has assisted in the positive identification of four antibiotics in the early stages of their isolations. Using characteristics of these four, in the form of crude isolates, the computer was able to pinpoint four antibiotics; actinomycins IV and V, griseorhodin A, and porfiromycin; as the most probable matches. The computer prediction was confirmed by appropriate chemical analysis. The other three products still unidentified do not match any compounds in the data base to the degree required to insure their identification. Analysis of the results of the searches have indicated their probable antibiotic type and this information has been used to specify additional isolation procedures.

M. BOSTIAN
K. MCNITT
A. ASZALOS
NIC, Frederick Cancer Research Center,
Frederick, Maryland 21701, U.S.A.

J. BERDY
Research Institute for Pharmaceutical
Chemistry, Budapest 4.
Szabadsagharcosok U. 47–49., Hungary

(Received April 2, 1977)

References

1) ASZALOS, A.; S. DAVIS & D. FROST: Classification of crude antibiotics by instant thin-layer chromatography (ITLC). J. Chromatogr. 37: 487~498, 1968
2) BETINA, V.: A systematic analysis of antibiotics using paper chromatography. J. Chromatogr. 15: 379~392, 1964
3) BERDY, J.: Recent developments of antibiotic research and classification of antibiotics according to chemical structure. Adv. in Applied Microbiol. 18: 309~406, 1974
4) BERDY, J.: International Center of Information on Antibiotics. Information Bulletin No. 10, p. 1, 1972
5) BOSTIAN, M.: A free format data base creation and search system. in preparation.