ASPECTS OF THE CBCC BIOLOGY CODE OF INTEREST TO CHEMISTS*

By ISAAC D. WELT, Ph.D., Director,
Cardiovascular Literature Project,** Division of Medical Sciences, National Academy of Sciences,
National Research Council, Washington, D.C.

The Chemistry Code developed and used by the Chemical-Biological Coordination Center has now been available in published form for a decade. This publication, entitled "A Method for Coding Chemicals for Correlation and Classification", has been and is still being profitably employed by numerous information groups in the chemical industry. In our own office, there is a file of some 65,000 compounds coded by means of the so-called "NRC Chemical Code" described in the above-mentioned publication.

Another of the intellectual products arising from that brave, premature venture into chemical-biological documentation, the CBCC, has now become available. Although it is a "posthumous" publication, many of the original contributors are still around, including the present author, to introduce the Code to its potential users.

The Detailed Biology Code with its accompanying Key is a large and impressive document, with a lengthy history. The present publication represents the 7th edition of a volume hitherto largely restricted to internal use by the CBCC staff and some selected outside groups. The original Code was devised by a number of NAS-NRC committees composed of subject-matter specialists with an interest in documentation. Extensive modifications gave rise to new editions. These changes resulted from experience gained in the actual use of the Code both for input and for the retrieval of chemical-biological data. As a result, the forthcoming publication consists of a system which has been tested thoroughly in practice. It is by no means a perfect approach but it can serve as a basis for further refinement and application.[1,2]

Since the main aim of the CBCC was to record, in great detail, the effects of all chemical substances upon all biological systems referred to in the literature, the Biology Code was constructed on a very broad basis. In order to permit continued expansion and growth, a "fixed field" format was chosen which utilized 70 columns of the standard 80-column IBM punchcard. In other words, a pre-determined number of columns were allotted for the codification of various descriptors, such as the "test organism," "target organ," "host," etc. To obtain maximum benefit from this arrangement, it frequently was found necessary to allot the same space to two or more descriptors which were mutually exclusive. Thus, if a "host" were not present, the space reserved for it could be pre-empted by a term for "environment." Unfortunately, such multiple assignments of valuable card space increased the number of rules for the use of the Code. Although the Code was designed to cover the entire broad spectrum of Biology, it is sufficiently refined to allow for the careful coding and retrieval of highly specific "bits" of information. Of special interest to the chemist are the fields devoted to the coding of the physical state of the applied chemical, the vehicle or solvent employed and quantitative information concerning the dosage level.

It is of historical interest to documentalists that the technique which came to be called "semantic factoring" was employed. One column was devoted to the coding of a number of verbs such as "increases," "stimulates," "enlarges," "decreases," "inhibits," etc. By simply adding such modifiers as "does not," negative effects were included. Each codable entry was thus able to convey a specific piece of information in the form of a concise sentence, often with dependent clauses.

Examples of actual coded material expressed in sentence form serve to clarify this point:

1. A lactic acid solution of the test chemical, injected intravenously into 125 gram male and female rats, at a dose level of 10-50 mg./kg. killed 50 per cent. of the animals. The LD-50 was calculated to be 21 mg./kg.

2. A single subcutaneous injection of 10 mg. of a test compound in 50 adult arthritic male patients, pre-treated with Atropine, increased the blood pressure of 45 of them within two hours.

3. An intramuscular injection of the test chemical administered to dogs 10 minutes before a subsequent intramuscular injection of ethyl alcohol decreased the peak blood level of alcohol by 50 per cent. within two hours.

4. The test compound, when added to a rat liver brei at a concentration of 0.5×10^{-5} M, produced a 50 per cent. inhibition of oxygen uptake when either formaldehyde or acetaldehyde was the substrate for the enzyme, liver aldehyde oxidase.

There is no "short-cut" to the intellectual effort involved in the selection of codable information from the pertinent literature. In fact, the variety and depth of coding possible with the CBCC Biology Code demands a meticulous analysis of the data. Hence, the Code can be used efficiently only by well-qualified information scientists who are constantly aware of the specific information needs of their organization.

On the other hand, the elaborate classification system provides an easy-to-use "thesaurus" or subject heading list in hierarchical form. Thus, in the "target organ" field, which consists of three columns, the first column may

*Presented before the Division of Chemical Literature, American Chemical Society, 138th National Meeting, New York, N.Y., September 13, 1960.
**Supported by Research Grant H-2045 from the National Heart Institute of the National Institutes of Health, U.S. Public Health Service.

be used for "heart," the second, for "ventricle" and the third, for "papillary muscle." Conventional methods of selection then can be employed for the retrieval of information concerning chemical effects upon the heart as a whole, upon the ventricle only, or upon such a specific structure as the papillary muscle.

The classification scheme, because of its broad scope, is of interest not only to the medicinal chemist and the biochemist, but also to agricultural and pesticide chemists. The adoption of this integrated system by groups engaged in the coding of chemical-biological data for rapid and efficient input and retrieval could provide at least a general framework, subject to extensive modification. That is, the Code is so constructed that it is not necessary to accept each division, subdivision or term in order to benefit from the system as a whole. Major revisions and reassignments of fields in order to better meet the more restricted interests of, for example, an information division in a drug company, are feasible. The classification system can be expanded and updated as necessary. Furthermore, although originally designed for use in conjunction with the well-known CBCC or NRC Chemistry Code, it is not unalterably wedded to this system. Most methods of chemical ciphering amenable to machine manipulation can be integrated with it to achieve a usable system for the retrieval of general or specific data concerning the relationships between chemical structure and biological activity. This goal, incidentally, was one of the major long-term aims of the Chemical-Biological Coordination Center.

This led to the development of several subsidiary classifications. One of these involved a hierarchical classification of enzymes. This makes it possible to obtain information on all "dehydrogenases," for example, or about the behavior of "lactic dehydrogenase" only. Unfortunately, in order to achieve publication of the Code, and due to the demise of the CBCC, the Enzyme Code is several years behind the times. However, it was so constructed that addition of the numerous new enzymes discovered within the past few years can be carried out easily without changing the classification framework. The use of suitable modifiers makes it possible to record either the effects of a test chemical upon an enzyme or enzyme system or the action of the enzyme system upon the test chemical and the products so obtained.

The importance of the metabolic alterations of the test compound produced by living organisms gave rise to a "chemical reactivity" code. That is, all chemical reactions known to occur to test chemicals administered to organisms were classified and assigned code symbols. For example, the process of "conjugation" is listed and coded by three alphanumeric characters. A fourth symbol further specifies the type of conjugation, such as "with glucuronic acid," "with glycine," etc. As a further example, "acetylation" is coded as a more specific type of "acylation." The metabolic product of the test compound pound is also retrievable in coded form on the basis of its structure.

Examples of searches for information of this character, which may be of interest to chemists, can be cited:

1. Which compounds of specified structural characteristics have been tested for their effects on the Glutamic-Aspartic Transaminase system? What were the results? If the system was inhibited, what compounds may counteract this inhibition?
2. Have any agents of chemical composition similar to certain newly-synthesized drugs been shown to undergo hydroxylation of the benzene ring in living organisms? If so, does this decrease their pharmacological effects? What are the effects, if any, on their systemic toxicity? Which enzymes are involved?

Since the entire field of Biology is covered by the Code, meaningful correlations can be made between different fields. For example, mechanism of action studies concerning organic phosphorus insecticides may be applicable to the study of the so-called "nerve gases" and "parasympathomimetic" drugs.

Before proceeding to a critical evaluation of some of the shortcomings of the CBCC Biology Code, it must be remembered that it was a pioneering effort, conceived in the very early days of scientific documentation. It often has been the experience of many of its early contributors and users, to learn of the independent "re-discovery" of some technique or approach which had been incorporated in the Code or in the Key which accompanies it and contains the directions for its most effective use. Unfortunately, this field of endeavor changed so rapidly, and experience accumulated which led to frequent Code revisions, that it was deemed advisable to delay publication until now. Despite these revisions, certain shortcomings remain as a result of its comparative antiquity. Designed long before the advent of even the IBM-101 Statistical Machine, not to speak of modern high-speed computers, the Biology Code is conceptually limited to the application of obsolescent or obsolete machine methods. For example, the 80-column limitation of an IBM punch-card gave rise to the multiple use of fixed fields. This, in turn, resulted in the rapid accumulation of numerous rules and regulations for the proper use of each field under a variety of circumstances. The inadequacies of the original instructions which became increasingly evident as the Code was put into practical application, resulted in further rules and sub-rules. These were collected in what is termed the Key to the Code. Overambitious planning with the view toward taking care of all coding problems not only at the time when they

arose but also in the forseeable future, resulted in a Key which is more bulky than the Code itself. Thus, as amendment is piled upon amendment, the original "constitution" becomes increasingly difficult to interpret and to use as an instrument of policy.

This is particularly true in the statistical or evaluation section of the Code, which I have not heretofore mentioned. In addition to coding the pertinent material, using quantitative data wherever available, it was decided to attempt comparative evaluations of the chemical action on a given biological system by means of acceptable statistical criteria, such as "probits," "chemotherapeutic index," etc. Space was provided on the punch-card for the recording of the results of such evaluation schemes. By and large, the time involved did not justify this intriguing approach. If thousands of compounds are tested by means of a small number of standardized tests, such as in screening programs, this aspect of the code might well be worth further trials. Otherwise, it merely complicates the efforts of the coder, already frustrated by an overwhelming number of coding rules involving the hurried turning of pages and scanning of possibly-pertinent footnotes.

It is sincerely hoped that with the availability of this publication, a new impetus will be given to progress in the coding and mechanical handling of biological data, leading to a greater degree of standardization.

Despite the many and serious unresolved problems which beset the art and science of handling large masses of purely chemical data, its biological counterpart, inherently more complex, is much more undeveloped. It is hoped that with the publication of the Detailed Biology Code of the Chemical-Biological Coordination Center, a strong impetus will be given to progress in the biological information area.

It should be mentioned in closing that the Code has been published recently by the National Academy of Sciences—National Research Council [3] and that this publication was made possible by generous support from the Office of Science Information of the National Science Foundation.

BIBLIOGRAPHY

[1]George A. Livingston and Isaac D. Welt, Advances in Documentation and Library Science, 2, 250-270 (1957), "Chemical Structures and Responses of Organisms to Applied Chemicals."

[2]Congdon G. Wood and Isaac D. Welt, Journal of Agricultural and Food Chemistry, 4, 886-888 (1956), "A Multi-indexed Machine-Sorted, Punch Card System for Pesticide Metabolism Data."

[3]Philip G. Seitner, Editor, "Biology Code of the Chemical–Biological Coordination Center. Key to the Biology Code of the Chemical–Biological Coordination Center," Publications 790 and 790K, National Academy of Sciences–National Research Council, 1960.