Storage and Manipulation of Data by Computers for Determinative Bacteriology

MICAH I. KRICEVSKY and LEWIS M. NORTON

National Institute of Dental Research and Division of Computer Research and Technology, Bethesda, Maryland 20014

Determinative bacteriology inherently involves acquisition and manipulation of many small pieces of information. Much of this information is codable directly in binary (i.e., "yes" or "no") form. Digital computers are uniquely suited to perform the desired data processing. Although at least one extensive coding system with some associated computer programs has been developed, much remains to be done before a universally applicable microbial data system is available. Some of these remaining computer technology tasks are the development of complete systems for data handling, data compression methods, and common specifications for data communication. Common specifications also are required for the methods for elucidating the bacterial characteristics which are to be encoded. Some classes of data cannot be coded reasonably in binary form. In such cases, different coding methods will have to be developed along with computer techniques to manipulate and analyze these classes. Some examples of these classes, each with their unique problems, are sources of isolation, nucleic acid hybridization, serological typing, and genetic mapping.

One of the basic problems in determinative bacteriology is to ascertain if a given isolate is a member of a defined population. This is usually accomplished by examining a subset of the characteristics of the isolate. The purpose is either to predict if an unmeasured characteristic is likely to be present (e.g., is the organism pathogenic for humans?) or to correlate information (e.g., occurrence of a given organism in parts of an ecosystem). As many biochemists or food technologists have discovered, it has not been necessary to identify an organism to study or use it (except for the "trivial" need at publication time). Measurement of genomic relatedness does not require identification. Deoxyribonucleic acid (DNA) hybridization studies can proceed quite happily without ever making an identification.

Identification is meaningful only when the allowable range of variation of a population is defined. Definitions of taxa are prepared by techniques which are either probabilistic or exact. Probabilistic techniques for preparing definitions of taxa are those in which only the estimate of relationship, and not the traits themselves, is considered (weighting may or may not be used, as it does not affect the concept). For illustration, a DNA homology study may be considered to be a probabilistic technique in that the gene products coded for by the DNA are not considered in determining the extent of the relationships involved among the strains under examination. Exact techniques for obtaining definitions are those in which a prescribed set of specific characteristics must be present (also termed monothetic taxonomy [9]). (The biochemist wishing to effect the oxidation of galactose is unlikely to be interested in a non-galactose-utilizing strain of *Pseudomonas putida* even though the isolate is only a one-step mutation of the type strain.) Combinations of probabilistic and exact definitions are possible and may be quite useful.

It is not within the purview of this essay to elaborate on how taxonomic descriptions are arrived at or to evaluate taxonomic methods. Rigorous treatments of the concepts of bacterial classification are available (9). In all methods for identifying bacteria there is a need to examine a large number of isolates to construct a reproducible, information-rich taxonomy.

If the purpose of determinative microbiology is to establish mechanisms for identifying new isolates, it would be useful to analyze some properties of microbial classifications in general. Traditionally, the taxonomist seeks to establish taxa, or categories, to which he can assign labels, each of which covers a set of isolates having a number of properties in
common. The numerical taxonomist differs from the "classical" taxonomist mainly in the approach used. The numerical taxonomist looks with statistical tools. The classical taxonomist uses inductive reasoning without the aid of a computer. Both approaches involve searching for linked, stable characteristics to use as criteria for membership in the taxon. Here, the individual properties are not the main focus, but rather stability and frequency of occurrence in the group.

A variety of classification schemes, designed to order strains in hierarchical categories, are widely used in microbiology. If an organism is described as an obligately aerobic chemolithotroph, a great deal of information about the organism is transmitted via the description. Microorganisms of medical interest may be classified by source of isolation, disease caused, susceptibility to antibiotics, etc. The food or dairy technologist may classify organisms by fermentation patterns and, thus, ability to cause spoilage or to produce a desirable product, such as cheese.

The first task in establishing a determinative system is obtaining, recording, and analyzing characteristics of large numbers of isolates. It is a basic precept, and a sound statistical statement, that the more isolates studied the better chance an unknown isolate will be correctly identified. The clinical laboratory technician describes this process as "experience." A number of mechanisms exist for circumventing the need for personal experience. Standards can be established by means-of-identification "keys," by cluster analysis, or by other conventions which have the net effect of delineating a minimum set of characters sufficient to include or exclude fresh isolates within a given group. Such standards imply that large numbers of characters and isolates have been studied and the results have been communicated.

Unfortunately, the history of such communication is one of arbitrariness born in isolation. Only recently have groups of microbiologists combined forces in working committees to develop common procedures. Within a committee, communication is, too often, by "brute force," i.e., hand-coded tables. Communication between committees takes on the flavor of international political negotiations over territorial rights and boundaries. Misunderstanding, incompatibility, and inconsistency in the analyses are common.

Computer technology offers a way out of these difficulties. Data in machine-readable form have the obvious advantage of being transmitted faster, cheaper, and more accurately than by traditional methods. The single biggest advantage provided by the use of computer technology is actually a function of the limitations of the computer, i.e., the inherent rigidity of machine processing. Because computers do not interpret data, rigid rules must be generated and adhered to for this kind of data handling. For success within a reasonable time, rules for the data processing must be universally applied. A new set of rules for each group of users or group of organisms defeats the purpose of using computers. In generating global rules, fundamental commonalities in describing bacteria are sought. Hopefully, a common "language" for communication of descriptions of isolates and logical groups of isolates can be generated. Thereby all laboratories can exchange their combined "experience" quickly and easily.

Widely disparate techniques, such as fluorescent antibody staining, carbohydrate fermentation patterns, and antibiotic sensitivities, share a common trait. They assess, at least in part, relatedness of isolates. The informational content of some tests can be great. Testing for the presence of a single protein by immunological methods can provide reliable identification. Alternatively, assessment of the trait "glucose utilization" has only low informational content since a variety of metabolic pathways exist by which bacteria can utilize glucose. In actual fact, it is the specificity of the question which is low, hence many questions of this type are required to obtain the same amount of information as, perhaps, the result of an antibody reaction.

If it is assumed that 100 to 400 bits of information are required to define a species (or to identify a given isolate as a member of a species), the problem can be examined in another way. A sequence of 100 to 400 amino acids in a particular protein can be assessed serologically or by electrophoresis. Alternatively, 100 to 400 proteins having particular functions, irrespective of the amino acid sequences accomplishing those functions, can be examined. It matters little theoretically. A great deal of redundant information in the bacterial genome presumably exists because many determinative strategies give similar answers. Thus, the numerical taxonomist, the classical taxonomist, and the molecular taxonomist rarely, if ever, arrive at radically different taxonomic conclusions. The exceptions comprise only a very small proportion of the total.

The informational content of a G+C ratio is
low, viewed in the determinative sense. The total number of slots or bins available for placement of strains are between 60 and 70 (assuming an overall precision of 1%). It would be extremely unlikely to obtain overall DNA base compositions greater than 80% or less than 25% G + C. Furthermore, most of the G + C data cluster about nodes rather than randomly distribute throughout the available space. Thus, this criterion is determinatively useful only in the negative sense. It tells us when an isolate is out of bounds.

With DNA homologies, the same problem of number of bins exists. In addition, the technique is not practical for widespread routine determinative use. Since this discussion is focused on computers as an aid in determinative bacteriology, we shall consider the area of generation of a global coding system for the description of DNA homologies. The problems in developing such a code are great. This topic is discussed later.

A start has been made at developing a comprehensive coding system for computer handling of descriptive data for the bacteria (7). It is in use in various institutions around the world such as the American Type Culture Collection, University of Maryland, and the National Institutes of Health in the United States, the National Collection for Type Cultures in Great Britain, and the University of Queensland in Australia. Programs interfacing with, and utilizing, this coding system are being developed by a number of laboratories.

The system provides the capability for coding, in computer-compatible form (mainly binary), a large array of biochemical, physiological, and morphological characteristics of bacteria. In addition, limited information of a survey type can also be coded on such items as source, time of day, date, and place of isolation, G + C content, etc. Forms or coding sheets have been developed for primary recording of the data. A standardized file structure for storage of the data on magnetic tape has been worked out in collaboration with Skerman and his co-workers. However, magnetic tape formats imposed by different computer manufacturers have left us with some formatting problems.

Query systems have been developed by the group in Brisbane (a batch system) and ourselves (an on-line interactive system). Such query systems have utility in developing taxonomic schemata and in making simple, non-automatic identifications. They have potential in that future expansion is possible. They can be interfaced with various automatic determinative strategies.

**PROBLEMS IN STANDARDIZATION**

In setting up a data bank, there are tasks to be undertaken, namely, standardization of parameters such as communication methods (e.g., magnetic tape characteristics), programs for creating and manipulating data files, and the tests themselves which are performed on the bacteria.

**Magnetic tape standardization.** The problem of standardization of magnetic tapes for communication is, by contrast, easy to solve. It only awaits an agreed-upon convention. Either the scientific community can adopt a particular tape configuration as a common standard (should it be 7- or 9-track; 256, 512, 800, or 1,600 bits per inch; ASCII or EBCDIC codes, etc.), or it can decide to adopt a format for the information on the tapes and to leave interconversion of tapes from configuration to configuration to a few centers capable of handling the task. A survey of the users' preferences and capabilities in this area should provide adequate information for making the decision.

**Data manipulation programs.** Care is needed in making decisions about standardization of computer programs for creating data files and for manipulating the files for a vast array of possible uses. To maximize the utility of a data bank, it is important to keep the system open to all investigators. The programs which create the initial general files are best standardized from the start. However, it would not be appropriate to standardize methods for numerical analysis, taxonomic cluster generation, determinative strategies, or report generation. Rather, standard programs should be created which allow the system to be interfaced with files of existing data from other laboratories or institutions. Such programs facilitate inclusion of data in the system that may not be recorded in the standard format of the data bank. The main sources of such data are (i) files of data of historical interest from the literature or laboratory records and (ii) data generated in computer-compatible form for a very specific use. Programs will have to be written to interface with output programs an individual user might wish to use. This could be done by generalized reformatting programs acting as translators between the system's data file structure and that of the target output programs. Users of translator programs need to specify the order, the nature of the characters used in one file, and their corresponding meaning in the target file. It should be understood that a one-to-one relationship among files is not necessary. For
example, a file for input might have a gram-positive straight rod coded as a "+." The translator might be told to interpret the "+" as 1,1,1, for items 2, 13, and 47 of the data file. An important versatility for output translation is provided by the ability to make various items of data equivalent (or to be combined by other Boolean algebraic operations) and to return a logical answer for manipulation by other programs.

The MYCSIS program prepared by a group of Canadian workers (8) provides a system for handling strain names and synonyms of bacteria, etc. Names in the Index Bergeyana (1) can be entered into the computer directly, rather than via alternatives such as numerical codes (e.g., the SNOP code [2] used in clinical work) or a check-list of choices. This is because name changes and the lists would be quite long as well as subject to error.

Detailed histories of strains should not be entered into the data bank unless the computer is to be used to search for information within the history. If the computer is to be used only to keep track of the history, it would be more practical to keep the information in a computer-indexed microfilm system. Perhaps synonymous strain numbers could be stored in the computer, but little else of the history of a strain that passes from laboratory to laboratory could be justified for the data bank. (If the microfilmed text were in a "standard" optical reader type font, the limitation would be minimal.)

Standardization of test methods. The first problem in standardization of tests and methods is that each worker may have a pet method that works well in his laboratory and for the bacterial group of interest to him. If this problem seems trivial, consider the number of Gram staining methods which have been published. Each of us is a charming fellow, quite adaptable to the changing needs of science. But what of that stubborn brute, your colleague in the laboratory down the hall—will he graciously agree that your method, which is superior, should be adopted as the standard method?

Another pitfall in test standardization is the confusion of a test with the character being scored or coded. In most cases, the attribute being assessed is quite independent of the method used in the process of assessment. Utilization of glucose by an organism can be assessed with equal validity by a variety of different procedures such as growth measurement, acid production, gas production and, most directly, measurement of the disappearance of glucose. An illustration of what can go wrong if this distinction between organism characteristic and test method is not appreciated is found in a common method for detecting arginine dihydrolase. A growth medium and pH indicator are used. The action of the arginine dihydrolase results in the formation of basic products. Thus, the indicator changes color. However, the test is also positive when arginine is metabolized by another known pathway, i.e., via primary attack on the arginine molecule by arginine decarboxylase. Possible solutions to this problem are to change the test (e.g., measure the ammonia resulting from the dihydrolase action) or to rephrase the attribute being scored (e.g., "Arginine utilization results in the formation of basic end products").

It would be counterproductive to make a separate question for each method of assessing a characteristic. For practical reasons, it may be useful to include questions about the results of commonly used tests to avoid losing data already recorded in this form. Perhaps what is needed even more than "standardization" of tests is a compendium of tests for which the validity has been ascertained and cross-indexed with other tests equally valid and equivalent. It would contain the kinds of organisms and the range of conditions for which the tests are valid. Special attention should be paid to finding general tests valid for the widest possible range of organisms and conditions. When a compendium is available from a fully constituted and recognized committee which has validated, interpreted, and chosen the tests, it would be a simple matter to develop a coding system for the tests in the compendium. This coding system would then be utilized in associating the test method used with the attribute of the organism scored in the data bank.

CODES FOR SOURCE OF ISOLATION

Codes for describing the source of isolation of a strain are required. A statement that a dental epidemiologist might code is: "The strain was isolated at the Forsyth Dental Clinic from a carious lesion, upper left second molar, buccal surface, at 1315 h, 16 May 1963." An aquatic microbiologist might wish to code the statement: "The organism was isolated from a small inlet in the Chesapeake Bay at 1315 h, 16 May 1963, from a depth of 2.5 m." Entering these statements, as such, into a computer system would be productive only if the computer is subsequently used to process the
information. The computer should not be used only as a horribly expensive printer to regurgitate the information, unchanged. It should be noted, also, that processing natural language by computer is an exciting, but complicated and expensive, endeavor. As yet, such computer techniques are not very successful. A further complication is the requirement to build lists of synonyms (e.g., isolated from little finger = fifth digit), as well as translations, among various languages.

To avoid these problems, a standard vocabulary list of descriptive words can be established. The user can select from these words, using them to enter data or to interrogate the system. The advantage lies with the "code" being easily understood, i.e., the words themselves are the code. The disadvantages are as follows: (i) the length of many of the words or phrases, (ii) the special computer programs required to process even a circumscribed set of words, and (iii) the impracticality of processing words in a variety of languages. To avoid these problems, one could develop a master list of words and phrases, as suggested above, but with a code number associated with each word or phrase. Communication with the computer would be accomplished by entering either the original word or phrase or the number. Internal processing would always be done by using the number. Simple encoding and decoding programs would take care of the translation if words or phrases are input or output. The language problem is thereby greatly simplified, since the words and phrases are not used internally in the system.

The types of information that describe the source of isolation are primarily geographical, anatomical, taxonomic, temporal, and association with other phenomena. Each of these types of information has different coding characteristics. Some are simple and some are complex.

Geographic and time descriptors are simple to code. By using latitude, longitude, and distance above or below sea level, any location on the planet can be coded. The same logic can be applied to the moon and planets, should an organism be isolated from a Mars canal. Presumably, it will be possible to adapt one of the existing computer geographic information systems from the space or meteorology programs. Numerous programs have been described for handling time information in computers.

Taxonomic codes are useful to record which plant; animal, or mineral (or other nonliving material) was the site of isolation. Such codes will probably be available in whole, or in part, from workers in relevant disciplines. Some additional space will have to be provided so that unusual features can be described (e.g., "brackish water" or "hybrid high-lysine corn"). For reasons stated previously, such natural language entries should be minimized.

Taxonomic codes require development of anatomical and geometric modifiers. It is not sufficient to describe a source as "Acer saccharinum; leaf; top." Ambiguity must be avoided. Rules of grammar have to be formulated to distinguish between the top of a leaf and a leaf at the top of a tree. Elements of the codes required are underway by taxonomists in the relevant disciplines. Presumably, these codes will be directly adaptable as modules to be used in a microbiological system.

For a variety of applications it is desirable to record and use data for phenomena associated with the strain. Such phenomena would be either linked to the source (e.g., "from an abscess" or "from ropy bovine milk") or to a special characteristic of the strain (e.g., "used in manufacture of dextran"); or "used in producing a vaccine", or is "a standard strain for checking antibiotic sensitivity").

The above codes will have to be supplemented with provision for short natural language descriptions to accommodate unanticipated situations. However, such techniques should be used judiciously. As with strain histories, most of the associated information is more properly stored in another form such as on computer-indexed microfilm.

DATA COMPRESSION NEEDS

As the amount of stored data increases, the problems of storage, manipulation, searching, and the like also increase. For example, records on antibiotic sensitivities of 18,000 strains isolated over a 4-year period at the National Institutes of Health Clinical Pathology Department fill a half reel of magnetic tape used for primary storage in the form described by Rogosa et al. (7). Clearly, a mountain of tape will result if nothing is done to increase the efficiency of storage and utilization. It would not be appropriate to discuss here the many stratagems available from computer and information science for internal data compression. Rather, some observations on ways of initial data handling will indicate the possibilities.

Much of the data to be submitted for storage and use in a large data system will be descriptions of isolates which, once studied, are
no longer available. In spite of the unavailability of the cultures, the data have to be recorded for subsequent processing in taxonomic, epidemiological, historical, and ecological studies. In most cases, a set of data will be submitted as a standardized, highly structured set. Further, it is likely that there will be great redundancy in the strains in a given set. For example, some strains are isolated with great frequency from clinical pathology specimens. Of the 18,000 isolates alluded to above, 18% were identified as *Escherichia coli*. That is, the isolates matched the criteria established for *E. coli*. However, the patterns of antibiotic sensitivity for this group of isolates varied considerably. Thus, the record need only be the species name of the organism, the date, the source of isolation, and the pattern of antibiotic sensitivities. By this means, the data would be compressed about 50%.

For automatic data compression, the redundancy occurring in large data sets can be utilized. The techniques of cluster analysis or other frequency-determining methods can be used to determine those strain descriptions occurring with greatest frequency. These modal descriptions would be recorded in toto. For the others, only the differences with respect to the nearest modal organism need be recorded. Since the modal organism is, in this case, generated from the frequency of isolation data, taxonomic significance need not be ascribed to the process. A theoretical median organism already has been applied to data sets.

Efficient utilization of data in a system will require special subfiles. The file organization should be designed so that answering the most frequently asked questions is rapid and simple. For example, ecologists or epidemiologists would use files presorted by geographical place of isolation.

**NONBINARY CODE NEEDS**

Perhaps the most difficult and complicated task to be faced is that of developing nonbinary codes. The terminology, in this case, refers to the concept that the data recorded are not in a form which can be reasonably coded by "yes" or "no." Instead, the property tested is linked to properties of other isolates (or mutants). Three examples, with some of the considerations appropriate to each, will illustrate this problem.

**DNA homology.** The most straightforward of the three examples is that of DNA homology data. Clearly, it is not sufficient to record data on a hybridization experiment without recording data for both members of the pair. In addition, different experimental conditions yield different results. DNA homology measurement is an estimate of relative homology, not absolute. Hence, the supporting data must be provided, i.e., strain used, conditions employed, and the like.

If the absolute homology could be recorded in a manner independent of the method, only then would it be sufficient to record the two isolates tested, the extent of hybridization, which macromolecules were hybridized, and any other parameters which were required to describe fully the relatedness of the pair.

If the methodology cannot be sufficiently standardized so that the data obtained by applying various techniques are directly comparable, then a much more detailed coding scheme must be developed which records the method as well as all the other parameters. However, as the amount of information needed for a full description of a single result increases, there is a loss in meaningful content.

Recording data such as homologies in a purely binary form would be an arduous task, at best. One would have to ask a new question for each test strain used. A way around this obstacle does exist. A set of bacterial strains can be chosen to represent the continuum of genomic relatedness for a given bacterial population. Precise relatedness, one strain to another, would be determined by an agreed-upon method (or even a group of methods for cross-referencing). A simple code then would be in the form: "Give the DNA homology of this strain with standard strain N" (where there is a separate question for each standard strain 1 through N). The user would select only those close to his strain or strains. Preliminary selection of the strains to test might be on the basis of G + C content. This method is analogous to constructing a genetic map for a single strain.

**Immunological tests.** A most powerful and widely used determinative tool in microbiology is that of the antibody reaction. There are many methods and classification schemes based on immune phenomena. A "brute-force" approach to coding antibody response would be to record (i) the strain of bacteria used in the immunization, (ii) the method of treatment of the immunizing strain, (iii) the species of animal being immunized, (iv) the method of immunization, (v) the method of preparing the isolate to be determined, (vi) the method of assessing the antigen-antibody reaction, and (vii) the result(s). All of this information is
required because neither antigens nor antibodies are pure or completely specific. Lack of specificity results in a profusion of cross-reactions. In turn, the exact spectrum of cross-reactions detected is dependent on the method of assessment.

Confusion is added by the technique of purifying an antiserum by adsorption with selected antigens. Although specificity of an antibody preparation is increased, the exact limits of the specificity are unclear. Thus, the prospect of an unwieldy, arbitrary, and open-ended list of coded items arises. Even more discouraging is that this type of information conveys no intrinsic knowledge about the organism. Within a selected, well-defined group of organisms, such techniques have excellent determinative utility. The Lancefield groupings for _Streptococcus_ and the Kaufmann-White scheme for _Salmonella_ are examples. In some instances, in these systems as well as others, such as the pneumococci capsular polysaccharide groupings, we see the beginnings of the ability to assign structural meaning to positive reactions. The problem centers around the arbitrariness and parochial nature of any given system of immunological grouping as currently constituted. In most cases it would be a hopeless task to try to make sense out of intergroup cross-reactions.

Distinction between asking questions about particular immunological reactions and questions about the presence or absence of specific antigens is necessary. The above considerations apply only to immunological tests. All questions involving the presence of specific antigens can be considered to be test independent. For example, if the antigen is a protein, electrophoretic or other techniques not involving immune reactions can be used for assessing presence of the component. Thus, a fundamental difference between the following two theoretical questions exists: (i) "The organism possesses the flagellar protein x"; (ii) "The precipitin test for flagellar protein x is positive." The first question is independent of the method of assessment. The second is not. It requires that both the specific protein (x) be present and the conditions of the test be correct and sufficient to detect that presence. Furthermore, assessment of the presence of protein x by some other method would not allow a positive answer to question 2.

Increasing utility of immunological technology can be achieved, allowing judgment of the genomic relatedness and/or the presence of specific antigenic structures. A system built around a series of carefully defined antigen-antibody pairs would be common to almost any stratagem. The antigens used to prepare antisera should be single molecular species. Further, it would be helpful if the immunizing antigens possessed some independently measurable structural or functional property. Examples of the use of enzymes as antigens for determining genomic relatedness are the work of Gasser and Gassar on lactic dehydrogenase of gram-positive lactic acid bacteria (3) and that of London and co-workers using multiple forms of malic enzyme, also in gram-positive lactic acid bacteria (4-6). When the determination of structural changes in the enzymes (as reflected in the immunological degree of relatedness) becomes known, coding systems based on such knowledge will be easily generated. A potential limitation of this approach is that a particular system chosen may be too restricted in scope. Since all cells must make de novo nucleic acids, some of the enzymes involved in these reactions might be applied more widely.

Another possibility is to invert the problem for coding purposes. Rather than study changes in antigens, it might be an advantage to characterize changes induced in the antibody. The knowledge of basic structures of antibody molecules, as well as changes in the structure brought about as the response to the introduction of antigens, is becoming available. A coding system can be based on the concept of recording the specific antibody species that reacts with the antigen being tested. Clearly, the structures of the antibody species will have to be predetermined.

Whatever the particular system chosen, it should have certain properties to be most useful. (i) It should be keyed to comparing structural properties of antigens. (ii) The antigen-antibody pairs should be chosen to have varying degrees of specificity, allowing determination of the relatedness of an isolate to groups of organisms of widely disparate dimensions. (iii) It should be based on antigens which are easily available (physically), such as surface antigens. (iv) For intergroup comparisons a series of antigen-antibody pairs having wide incidence should be available (such as nucleic acid metabolic enzymes). Whatever the coding schemes used, they must be able to both differentiate closely related strains and describe relationships of dissimilar organisms.

**Mutant description.** Coding of information for bacterial mutants provides another challenge. The types of information that can be coded are well defined. The problems would seem to revolve around deciding the point of diminishing returns for the effort expended in
designing and using a coding system.

In the coding system alluded to previously (7), the simplest level of coding has been implemented. One can record that a strain is a mutant of another strain. Phenotypic characteristics of each of the strains also can be entered in the computer file. Differences between strains thus represent the mutated gene or genes. The questions can be expanded to include operator genes, sexual genes, etc.

Not yet included is a method for recording distances within and between genes. Presumably it will be possible to devise a coding system which is independent of the method by which the mapping was done. For many of the potential tasks of the computer, relative distances from an arbitrarily selected starting point (in a positive or negative direction) should suffice.

Development of a universally applicable coding system to handle information at the level of complexity presently known about E. coli would find only limited utility for a considerable time. Thus, it is doubtful that such a system should be included in a general-purpose data system for another decade.

SUMMARY

Some general observations on the future outlook for computer-recorded data to be used for determinative and other applications can be offered. Perhaps the single most important consideration is to decide, on a rational basis, when it makes sense to use or not to use a computer. Reproducing data without any processing (as in a query such as "give me the history of strain number xxx") is a trivial use of a very expensive and complex resource. If this is all the computer is called upon to do with the information, the effort required for the machine to perform this kind of limited task is unjustified. Yet, if you have a system which does more sophisticated tasks, it will of course be able to do trivial ones, and may well be often used to do so. Further, it is wasteful to devise and to program systems for processing complex data of limited interest. Parochial uses are more properly left to the workers in the fields involved. As a more general utility appears, the parochial system can always be meshed with the general system.

The promise of computers for processing bacterial data is only now being appreciated. As the community of bacteriologists organizes itself in official and unofficial efforts, developments will come at an increasing pace. The systems will become increasingly useful as more people involve themselves in the development of a world-wide computer data system.

LITERATURE CITED