sponding structure and to compare different structures by using the CCT.

The CCT entries are ordered by using the semantic tree. The groups of entries corresponding to unsaturations, heteroatoms, and functional groups on one parent structure are sorted by locant order and output directly after the CCT entries for the parent structure represented by that semantic tree node. Each substituent, which may have subsubstituents as indicated by the semantic tree hierarchy, is then sorted first by locant order and then, for entries with the same locant, on the TIPE field. This brings substituents with the same locant and type (chain, rings, atom) together, and these are finally sorted on the SIZE field to order them from smallest first to largest last.

DISCUSSION AND CONCLUSIONS

This paper has described the processing carried out by the parser software in analyzing the syntax and semantics of IUPAC systematic nomenclature for some classes of organic compounds. The software has been successfully implemented in Turbo-Pascal and used within the nomenclature to structure diagram translator, which runs on an IBM PC-XT or compatible microcomputer. It confirms that the grammar-based approach using techniques developed for processing computer-programming languages can be applied to other less artificial languages. Two modifications were necessary, though, to these techniques. Most importantly, backtracking had to be introduced in the syntax analysis phase. Second, reference to a dictionary of valid morphemes was essential in the lexical analysis phase because of the lack of delimiters in chemical nomenclature.

In part 2,[2] a comment was made about the difficulty of developing a formal grammar from nomenclature rules as described in the Blue Book.[8] In this paper a good illustration of the complicated processing necessary to implement the current rules has been given. The rules for the alphabetic ordering of substituents are not only complex but also inconsistent as to which characters alphabetic ordering is applied. This depends upon whether the substituent is simple or complex. It necessitated the implementation of a data structure specifically for this one task, with code to detect simple and complex substituents and perform different processing on the two types.

ACKNOWLEDGMENT

We gratefully acknowledge funding of this research by the Laboratory of the Government Chemist.

REFERENCES AND NOTES

(1) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 1. Background and Introduction. *J. Chem. Inf. Comput. Sci.* (first of three papers in this issue).
(2) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 2. Development of a Formal Grammar. *J. Chem. Inf. Comput. Sci.* (second of three papers in this issue).
(3) Aho, A. V.; Ullman, J. D. *Principles of Compiler Design*; Addison-Wesley: Reading, MA, 1977.
(4) Thomas, M. I. *A Basic SLR Parser Generator*; Report No. 80/1; Department of Computer Studies, University of Hull: Hull, England, 1980.
(5) Lynch, M. F.; Harrison, J. M.; Town, W. G.; Ash, J. E. *Computer Handling of Chemical Structural Information*; MacDonald-American Elsevier: London, 1971; Chapter 6.
(6) Rayner, J. D. A Concise Connection Table Based on Systematic Nomenclatural Terms. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 108-111.
(7) See ref 2 Appendix.
(8) International Union of Pure and Applied Chemistry. *Nomenclature of Organic Chemistry, Sections A-F and H*; Pergamon, Oxford, U.K., 1979.

Canadian Scientific Numeric Database Service†

GORDON H. WOOD,* JOHN R. RODGERS, and S. ROGER GOUGH

Canada Institute for Scientific and Technical Information, National Research Council of Canada, Montreal Road, Ottawa, Canada K1A 0S2

Received November 14, 1988

Modern computer systems and telecommunications networks are being harnessed in increasingly innovative ways to deliver evaluated scientific/technical numeric data to the desk or laboratory bench. As an example of this development, the Canadian Scientific Numeric Database Service (CAN/SND) is described. CAN/SND provides international online access to factual databases in crystallography, molecular biology, spectroscopy, and chemical thermodynamics. In addition, CAN/SND carries out research in data storage, retrieval, and analysis techniques. The paper gives a description of the databases currently available, examples showing the variety of scientific questions that can be answered, and an outline of plans for virtually linking related databases for interdisciplinary searching.

I. INTRODUCTION

This paper describes the Canadian Scientific Numeric Database Service (CAN/SND), what it is, what it offers, and where it plans to go. All readers are almost certainly familiar with the computer as a tool for performing calculations and automating measurements; many are aware that computers may be used to search large bibliographic databases. Probably relatively few think of the computer as a means of accessing evaluated scientific data from the international literature. It is this latter use that will be highlighted here.

Immediately following, section II defines some basic terminology and gives some perspective on scientific numeric database systems. Section III reviews the background to CAN/SND and the databases currently offered. Section IV

† Presented at the Herman Skolnik Award Symposium on Scientific Numerical Databases—Present and Future, sponsored by the Division of Chemical Information of the American Chemical Society at the Third Chemical Congress of North America, Toronto, Canada, June 7, 1988.

* Author to whom correspondence should be addressed.

0095-2338/89/1629-0118$01.50/0 Published 1989 by the American Chemical Society

illustrates some examples of the questions these databases can answer, and section V is a survey of some plans for development.

II. DEFINITIONS

Databases may conveniently be divided into two broad categories: reference and source.[1] "Reference" includes those databases that contain bibliographic citations or references to other information sources; "source" includes those databases that contain numeric, textual-numeric, full-text or image information. In other words, a "hit" in a reference database points to a place where the desired information may be found, whereas a "hit" in a source database contains the desired information itself.

A scientific numeric database (SND) may therefore be defined as a source database, primarily in the textual-numeric subclassification. More specifically, it is an ordered collection of numbers whose values (1) correspond to various properties, parameters, or attributes of elements, substances, or systems and (2) are critically evaluated by experts prior to their being included in the database. For this paper, it is understood that the databases are machine readable.

Good scientific numeric databases are clearly much more than mere compilations of numbers. The important, and expensive, functions of review, evaluation, and correction serve to make the data more reliable than those found in the open literature and more useful because of the rationalization of factors like uncertainty statements and units of measurement.

To avoid confusion, the term scientific numeric database *system* (SNDS) is used here to describe a set of one or more scientific numeric databases combined with a suite of computer programs that enable the scientist or engineer to search the database(s), retrieve items of interest, and manipulate those items in various ways. Using a SNDS is, therefore, much more than simply thumbing electronically through a handbook to find a given entry, as the examples given in section IV will illustrate.

Some feeling for the relative abundance of scientific numeric databases may be gained from inspection of Figure 1.[2] Thus, in the spring of 1988, about 10% of all databases could be classified as SND in the sense defined above.

III. OVERVIEW

A. Historical Background. The Canadian Scientific Numeric Database Service is provided by the Canada Institute for Scientific and Technical Information (CISTI), which is a division of the National Research Council of Canada (NRCC), a corporate agent of the Government of Canada. Although the primary objective of CAN/SND is to make scientific numeric databases readily available in Canada, modern telecommunications now make it technically and economically feasible to offer the service internationally.

Since its beginning in 1980, CAN/SND has emphasized the dissemination and production of databases as well as doing research into data storage, retrieval, and analysis techniques. Online service, which began in Canada in 1981, became international in 1984.

From inception, the guiding philosophy of CAN/SND has been the importance of the scientist or engineer as end-user. Thus, decisions about enhancements and changes are user-driven rather than computer-science-driven. User feedback is actively solicited; changes must make the system more useful scientifically, not just more elegant or sophisticated.

B. Dissemination Methods. Interactive online access is the primary means used to disseminate the SNDs. Other means include (1) leasing tape copies of certain databases to those having adequate computing facilities and (2) providing a customized search service wherein scientists may submit a query for processing by CAN/SND staff. Those benefiting from the latter service are those who do not have access to a terminal or local computing facilities or those who need the system so infrequently that the effort required to effectively relearn the access protocols and search procedures each time a search is required represents an unacceptable overhead.

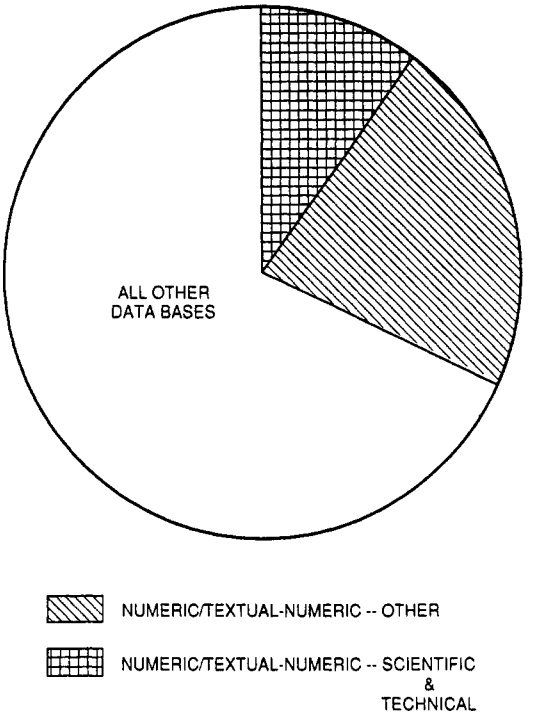


Figure 1. Ratio of scientific numeric to other database types.

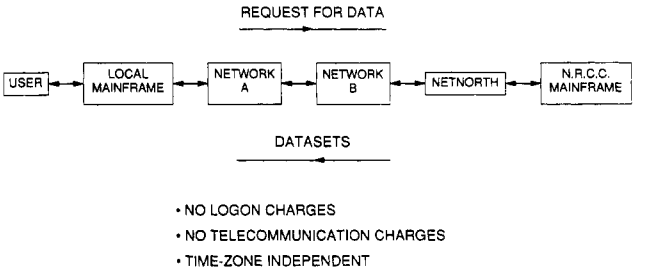


Figure 2. "Near online" schematic.

"Near-online", a recently offered variation on online service, is sketched in Figure 2. This access mechanism exploits the file-forwarding facilities of Netnorth, BITNET, etc. and a file-server on the NRCC mainframe. The file-server, a virtual machine operating under IBM VM/SP, is maintained in a so-called "sleep" mode from which it awakens upon the receipt of mail. Requests from accredited users, sent electronically via Netnorth and its associated and interlinked networks, are processed upon receipt, and the results are returned via Netnorth to the user. The machine then returns to its sleeping state until another request is received. Thus, without doing more than connecting to the local node of the network in his/her area, a user may have customized components of selected databases delivered to his/her mainframe within the hour. The obvious advantages of near-online are immediate, around-the-clock access and, essentially, the elimination of connect time and telecommunications charges.

C. Databases Available. Figure 3, a schematic of the current system operated by CAN/SND, shows the databases grouped in three basic disciplines: analysis, molecular structure, and molecular biology.

1. Analysis. SPIR (Search Program for Infrared Spectra)[3] is a collection of about 140 000 infrared spectra of some 96 000

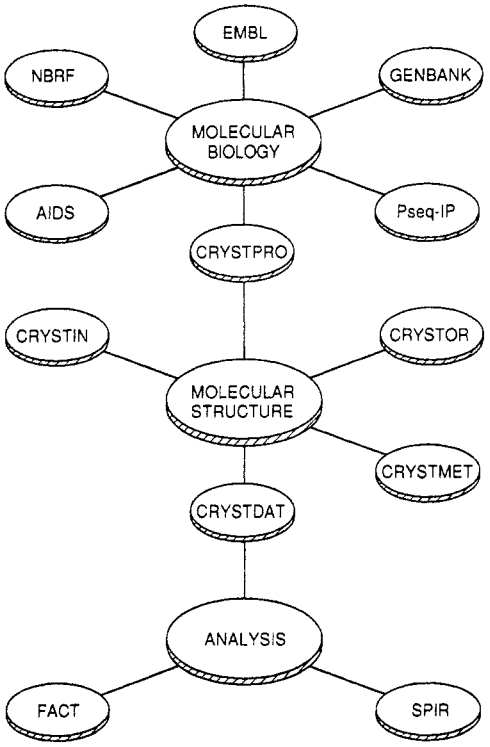


Figure 3. Databases available on CAN/SND (acronyms given in text).

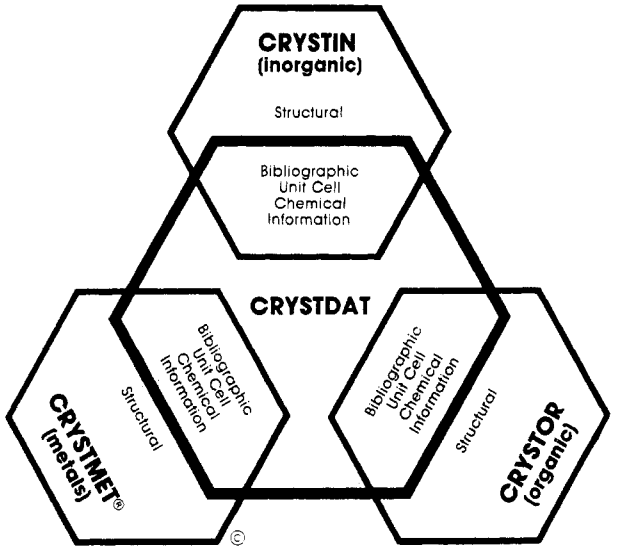


Figure 4. Relationship of CRYSTDAT to the other crystallographic databases.

compounds; entries consist of peak locations and some intensity information. (Compounds with spectra appearing more than once reflect different measurement conditions or workers.)

FACT (Facility for Analysis of Chemical Thermodynamics)[4] contains data on over 4000 inorganic stoichiometric compounds and aqueous organic systems along with a number of thermodynamic modeling and analysis programs.

2. Molecular Structure. Databases in this group feature structural data for crystalline solids. In addition, they contain related bibliographic, unit cell, and chemical information as appropriate. Figure 4 shows how the databases are related.

CRYSTDAT (NBS Crystal Data)[5] contains information on all crystalline solids for which the basic cell parameters are known. With about 130 000 compounds at present, CRYSTDAT may be employed to identify unknown substances using a minimum of crystallographic information.

CRYSTMET (NRC Metals Crystallographic Database)[6] contains information on some 23 000 metallic phases.

```
F*A*C*T SYSTEM, MONTREAL, QUEBEC, CANADA
COPYRIGHT 1982 THERMFACT LTD/LTEE
W.T. THOMPSON, A.D. PELTON, C.W. BALE

******ENTER A PROGRAM NAME OR PRESS RETURN******
******TO EXAMINE F*A*C*T LIBRARY ******

--> :reaction

CHEMICAL REACTION PATH CALCULATION (30 JAN. 84)

YOU WISH A 132 COLUMN (1) OR AN 80 COLUMN OUTPUT (2) ?
--> :2
ENERGY IN JOULES (1) OR CALORIES (2)?
--> :1
******ENTER EQUATION******
--> : 2 Ni     +     O2     =     2 Ni*O
SUBSCRIPTS
: (t,1,s2)     (t,p,g,)     (t,1,s3)
PRESS 'RETURN' WHEN READY FOR TABULAR OUTPUT
:
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
```

Figure 5. Query 1, invoking FACT to solve thermodynamic problems.

```
  2 NI          +     O2          =     2 NI*O
(T,1,S2)          (T,P,G)           (T,1,S3)

CALCULATIONS ARE BASED ON THE INDICATED NUMBER OF GRAM MOLES

     (T)        (P)        Δ H        Δ G        Δ V        Δ S        Δ U
     (K)       (ATM)       (J)        (J)        (L)       (J/K)      (J)
-------------------------------------------------------------------------------
--> :1000
    1000.0   0.200E-15   -471541.5      0.0   -0.411E+18   -471.542   -463228.6
--> :1250
    1250.0   0.164E-10   -468865.0      0.0   -0.627E+13   -375.092   -458473.8
--> :1500
    1500.0   0.295E-07   -466052.1      0.0   -0.417E+10   -310.701   -453582.4
--> : *
    1375.6   1.0E-9      -467473.6      0.0   -0.113E+12   -339.836   -456038.4
:/d
```

Figure 6. Query 1, results obtained from FACT.

CRYSTOR (Cambridge Structural Database)[7] contains information on about 67 000 organic and organometallic compounds.

CRYSTPRO (Brookhaven Protein Data Bank)[8] contains information on about 450 biological macromolecules, e.g., proteins, nucleic acids, viruses, and polysaccharides.

CRYSTIN (Inorganic Crystal Structure Database)[9] contains information on about 23 000 inorganic substances.

3. Molecular Biology. Databases in this family offer, in addition to the sequence data listed for each, bibliographic information and technical comments.

Pseq-IP (Protein Sequences—Institut Pasteur)[10] contains sequence listings for 10 000 proteins.

GenBank (Genetic Sequence Data Bank)[11] contains sequence listings for 18 000 nucleic acids.

EMBL (European Molecular Biology Data Library)[11] contains sequence listings for 18 000 nucleic acids.

NBRF (National Biomedical Research Foundation)[11] contains sequence listings for 8600 proteins and 2800 nucleic acids.

AIDS (Los Alamos National Laboratory)[11] contains sequence listings for the amino acids and nucleotides of the AIDS HI viruses.

SWISSPROT (Swiss Protein Sequence Database)[11] contains sequence listings for 6800 proteins.

Not shown on the diagram are the numerous associated search, retrieval, and analysis software packages available on the system. Information on these programs is available from the authors.

IV. APPLICATION EXAMPLES

It is beyond the scope of this paper to give detailed illustrations showing how each database may be exploited. The examples shown were selected to indicate the kinds of research that can be conducted and are not exhaustive either in detail or in scope.

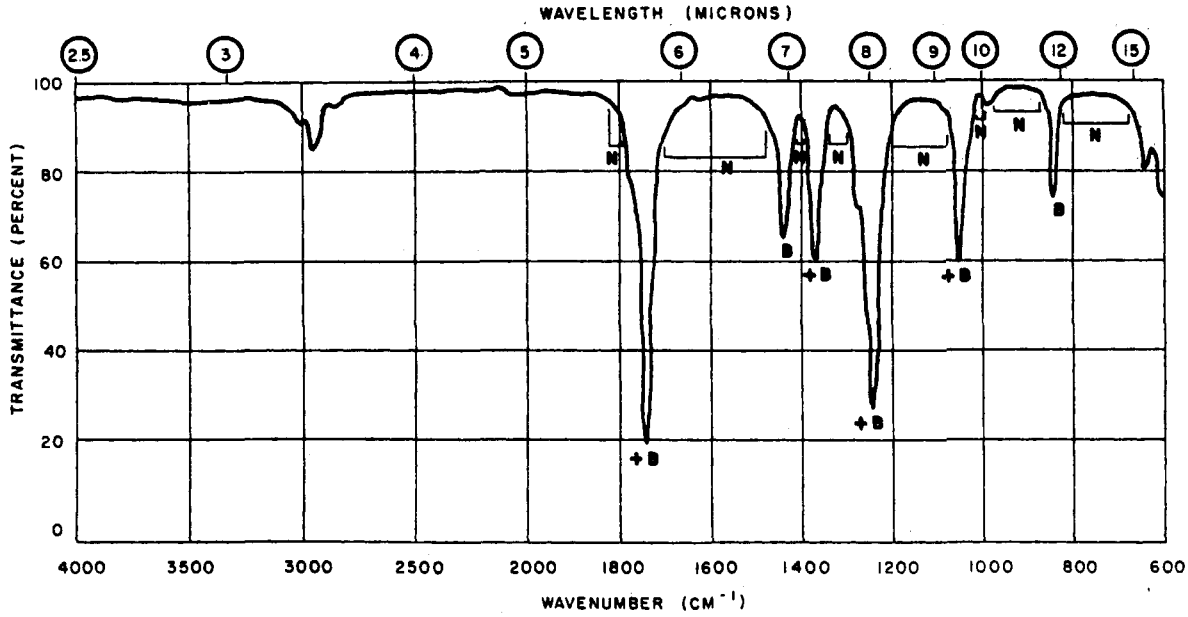


Figure 7. Query 2, infrared spectrum marked for input to SPIR.

```
> Find ele xla. or. xsc. and. xae. and. cu. and. o and spno 123.to.142

--Set 1 created with 7 hits
        :
ID:   808029
RC:   a=5.52 b=5.52 c=11.72 al=90.0 ga=90.0
CD:   sys=tetragonal spgr(CD)=P4/mmm spno=123 den=6.6(g/cc) z=1
EM:   Ba3 Cu6 La3 O14.10
FO:   La3 Ba3 Cu6 O14.10
NM:   Lanthanum barium cuprate
AC:   a=5.5253 c=11.721 spgr (A)-P4/mmm
RF:   J. Solid State Chem., 37, 151, 1981
```

Figure 8. Query 3, using CRYSTDAT to search for potential high-temperature superconductors.

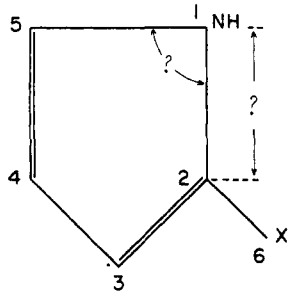
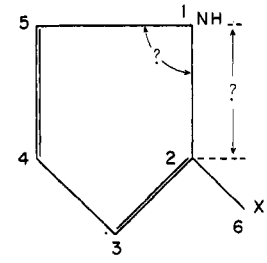


Figure 9. Query 4, formulation of the pyrrole problem.




```
AT 1   N   2   E
AT2    C   3
AT3    C   2
AT4    C   2
AT5    C   2   E
AT6    AA  1
BO     1   2   1   C
BO     1   5   1   C
BO     2   3   2   C
BO     2   6
BO     3   4   1   C
BO     4   5   2   C
NOLN
        :
DEF C-N        1   2
DEF ANGLE      2   1   5
```

Figure 10. Query 4, using CRYSTOR to solve the pyrrole problem.

A. Analysis

QUERY 1. What is the partial pressure of oxygen (O_2) in equilibrium with Ni–NiO mixtures at various temperatures; what is the temperature at which the equilibrium pressure of O_2 is 10^{-9} atm?

Figures 5 and 6, where user input is highlighted with an arrow, illustrate the use of FACT to address this question. In quite natural notation, the user is able to enter the reaction and specify the conditions of the variables: temperature (t), pressure (p), and state (gas, solid, etc.). To answer the first part of the question, the user enters some temperatures of interest, and FACT responds with the corresponding equilibrium pressures as well as the other thermodynamic parameters of possible interest. To answer the second part of the question, it is necessary only to specify the desired pressure, and the temperature is then treated as a dependent variable.

QUERY 2. The infrared spectrum shown in Figure 7 has been obtained from a substance whose identity is unknown. What compounds have similar spectra?

Before going online to use SPIR for this problem, the user would first manually scan the spectrum for significant peaks and mark them as shown, where "+B" implies a "strong" band, "B" implies a significant but weaker band, and "N" denotes regions where there are no bands. Once online, the user enters these data which SPIR uses as a template to compare with its spectra of known compounds. The result of a search is a listing of the 20 spectra most closely matching that of the unknown, ranked according to the quality of fit.

B. Molecular Structure

QUERY 3. What compounds are chemically and structurally related to known high-temperature superconductors of the form ABCuO in the space group $P4/mmm$ (where A is either a lanthanide or group III transition element and B is an alkaline earth)?

This problem, of great practical interest for those interested in synthesizing potential high-temperature superconductors, may be easily addressed by using CRYSTDAT as illustrated in Figure 8. Following the ">" prompt, the user requests all compounds having one element from the lanthanide group (xla) or the scandium group (xsc) and one element from the alkaline earth group (xae) as well as copper and oxygen; the user further specifies that any compounds satisfying these criteria must also belong to the space-group numbers 123–142, which correspond to the point group 4/mmm. Processing the query results in seven hits, one of which is shown as typical. With that information, the researcher has some potentially

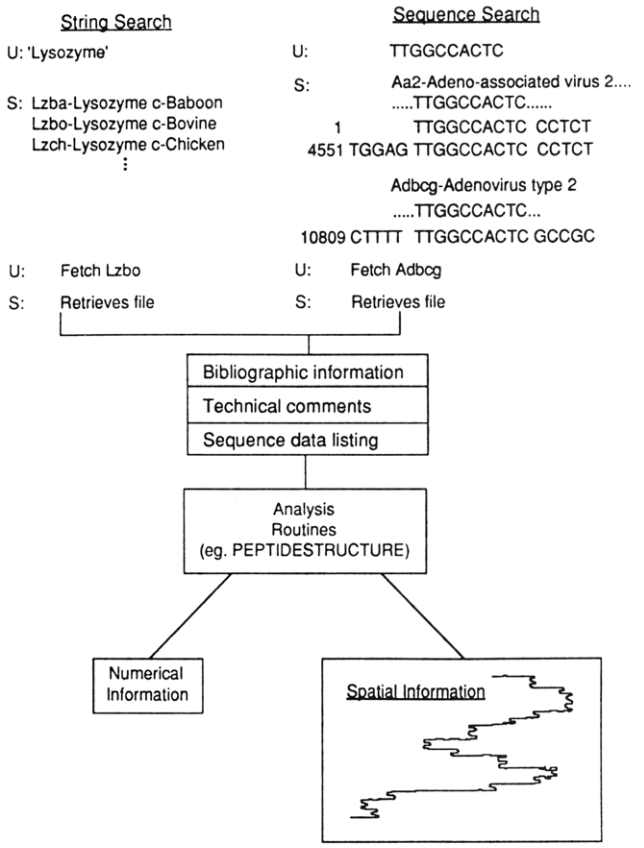


Figure 11. Query 5, string and sequence searches in molecular biology.

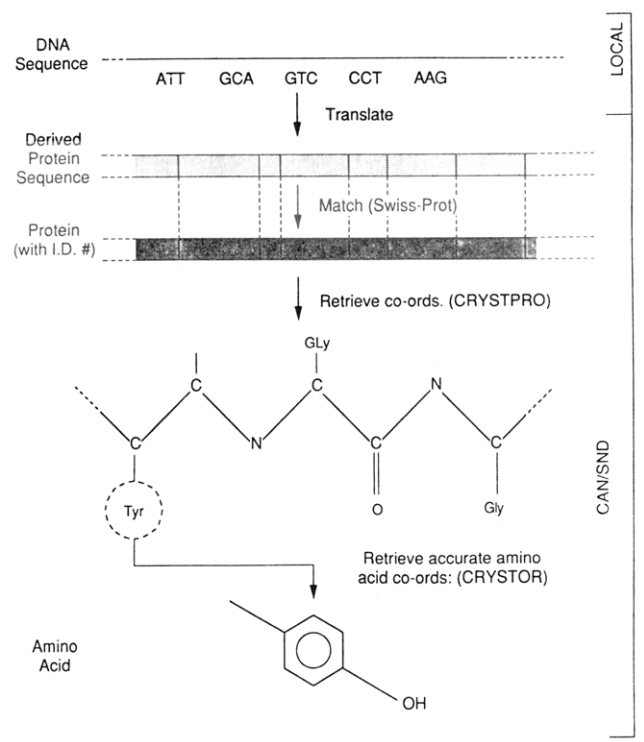


Figure 12. Schematic of an integrated molecular biology database system.

fruitful paths to explore in the quest for materials with the desired properties.

QUERY 4. Data are needed on as many compounds as possible in which a pyrrole (C_4H_5N) ring appears. Any ligand (X) may be substituted for one of the hydrogen atoms attached to one of the carbon atoms 2 or 5 (see Figure 9). In particular, how do the carbon–nitrogen bond lengths and the carbon (2)–nitrogen–carbon(5) angle vary with ligand X?

Using CRYSTOR (Figure 10), the user conducts a connectivity search describing the atoms and bonding arrangements in a systematic manner as illustrated on the left half of the diagram. CRYSTOR uses this description as a template to retrieve data for all the compounds having a pyrrole fragment satisfying these criteria. The numeric data associated with these compounds may then be passed, without further keyboarding, to internal programs to execute the required geometric analyses and to generate graphical output.

C. Molecular Biology

QUERY 5. What entries exist dealing with the enzyme lysozyme? In which entries does the nucleic acid sequence fragment TTGGCCACTC appear?

The approach to answering both parts of the question is illustrated in Figure 11. To look for lysozyme, an example of a string search, the user (U) enters the target string to which the system (S) responds with a list of entries containing the enzyme of interest. The user selects the one coded "lzbo" and requests the appropriate file. The search for the sequence fragment (right half of Figure 11) is trivial to initiate, but the response is profound. The system searches millions of nucleic acid patterns and reports not only the entries containing the fragment of interest but also the location and environment of that fragment wherever it occurs. Information from the associated sequence data listing may then be passed to internal analysis routines to derive numerical and spatial information.

D. Review

Generalizing from this set of examples, it is possible now to sketch the range of functions that scientific numeric database systems can perform:

(1) SNDSs retrieve items quickly, exhaustively, and accurately from large collections of data, retrieve along lines of thought for which the database compilers could not have foreseen the need for an index, and retrieve types of information too detailed and tedious for the human mind to handle readily (e.g., connectivity and sequence searches).

(2) SNDSs manipulate and analyze the data in a variety of ways.

(3) SNDSs simulate experiments with mathematical models, exploring processes like chemical reactions theoretically, thereby obviating the need to perform actual experiments or build prototype equipment.

(4) SNDSs formulate new ideas from observations and statistical inferences on the data themselves. (For example, the Cambridge Structural Database is a large body of reliable, basic data that has been used to gain information on the effects of substituents on chemical reactivity, on molecular flexibility, and on intermolecular forces.[12-15])

Obviously this list is not exhaustive; time and experience will reveal new uses. The ultimate power of a SND system is limited, not by the databases and searching algorithms available but by the imagination and creativity of those using them.

V. FUTURE DIRECTIONS—INTEGRATED SEARCHING

A. General. In the conventional use of a database system such as that of Figure 3, one tends to address queries to one or more of the databases on the basis of foreknowledge of what those databases contain. A facility for interdisciplinary searching would exploit the fact that these databases are mounted at one center and would lessen the need for a user to be familiar with the contents and attributes of each database because the system would automatically interrogate the databases appropriate to the user's query. To the user, the SND system would thus appear as an integrated whole with the distinctions between different databases tending to become

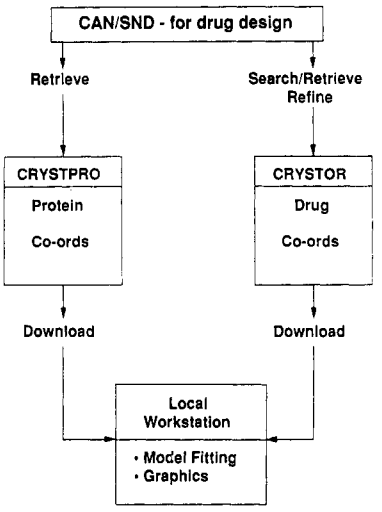


Figure 13. Information for drug design from an integrated database system.

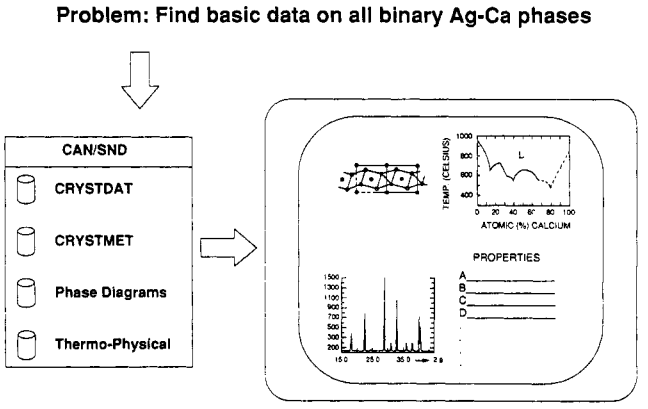


Figure 14. Basic property data from an integrated database system.

blurred. Concomitantly, users would gain confidence that possibly valuable information had not been missed because of their incomplete awareness of database specifics.

The examples that follow indicate the kinds of innovations that are planned in this area.

B. Molecular Biology. Figure 12 illustrates a situation where one has a DNA sequence, perhaps derived in the laboratory, and wishes to obtain the crystallographic coordinates of the related protein. Currently, it would be necessary (1) to derive the related protein sequence with a system program, (2) to match that protein sequence against the appropriate database (e.g., SWISSPROT) to obtain the protein identification number, and (3) to retrieve the coordinates for that protein from the CRYSTPRO database. Should more accurate coordinates for the amino acids be needed, the worker would have to use CRYSTOR. With the proposed enhancements it is planned to automate these steps, providing obvious savings in time and effort to the user.

Another instance is illustrated in Figure 13. Currently, someone wishing to design drugs would (1) use CRYSTPRO to look up the identification number of the relevant protein, retrieve the protein coordinates, and download them to the workstation and (2) use CRYSTOR to locate the relevant drug molecule and transfer the coordinates to the workstation. With the integrations planned, most of the steps would be automated, and the user would need connect to one inquiry system only rather than two databases.

C. Physical Properties. There are obvious advantages to one-stop shopping for physical property data. The scientist or engineer needing basic data on, for example, all binary silver–calcium phases (Figure 14) would currently need to search at least four different databases or combinations of databases and handbooks. With the enhanced system, the user need only specify the generic inquiry and the system would compile and present the data from the relevant databases in the most appropriate form.

VI. CONCLUSIONS

The Canadian Scientific Numeric Database Service (CAN/SND) is a source database system designed to meet the needs of scientists and engineers. The discussion and examples show that these SNDs are more than just a collection of bare databases that one searches electronically by intersecting various indexes; they are analysis and modeling tools with which one can answer problems that are intractable or impossible to solve by manual means.

REFERENCES

(1) *Directory of Online Databases*; Cuadra/Elsevier: New York, 1988; Vol. 9.
(2) Dewar, D. Private communication (based on online search of *Directory of Online Databases* ORBIT Search Service, May 1988).
(3) CAN/SND, CISTI, National Research Council of Canada, Montreal Road, Ottawa, Canada K1A 0S2.
(4) Dr. C. W. Bale, Ecole Polytechnique, P.O. Box 6079, Station A, Montreal, Quebec, Canada H3C 3A7.
(5) Mighell, A. D.; Stalick, J. K.; Himes, V. L. In *Crystallographic Databases*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; International Union of Crystallography: Chester, U.K., 1987; p 134.
(6) Rodgers, J. R.; Wood, G. H. *Ibid.*; p 96.
(7) Bellard, S.; Allen, F. H.; Kennard, O. *Ibid.*; p 32.
(8) Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. *Ibid.*; p 107.
(9) Bergerhoff, G.; Brown, I. D. *Ibid.*; p 77.
(10) Claverie, J.-M.; Bricault, L. *Proteins: Struct., Funct., Genet.* **1986**, *1*, 60–65.
(11) Beyman, R., Modelevsky, J., Roberts, R., Söll, D., Eds. *Nucleic Acids Res.* **1988**, *16(5)*, Part A.
(12) Taylor, R.; Kennard, O. Crystallographic Evidence for the Existence of C–H···O, C–H···N, and C–H···Cl Hydrogen Bonds. *J. Am. Chem. Soc.* **1982**, *104*, 5063–5070.
(13) Taylor, R.; Kennard, O. Cambridge Crystallographic Data Centre. 7. Estimating Average Molecular Dimensions from the Cambridge Structural Database. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 28–32.
(14) Bye, E.; Schweizer, W. B.; Dunitz, J. D. Chemical Reaction Paths. 8. Stereoisomerization Path for Triphenylphosphine Oxide and Related Molecules: Indirect Observation of the Structure of the Transition State. *J. Am. Chem. Soc.* **1982**, *104*, 5893–5898.
(15) Allen, F. H.; Kennard, O.; Taylor, R. Systematic Analysis of Structural Data as a Research Technique in Organic Chemistry. *Acc. Chem. Res.* **1983**, *16*, 146–153.