Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

In This Issue:

Departments

News Briefs

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

Articles

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

Dear Reader:

I have been appointed Chief Editor of the Journal of Research of the National Institute of Standards and Technology so that Karl Kessler may devote full time to his principal job of serving as the NIST Associate Director for International and Academic Affairs. It is my intention to continue the sound policy, so ably established and implemented by my predecessor, of bringing to the readers of the *Journal* significant NIST archival articles on new measurement results, methods, and instrumentation in the physical and chemical sciences and in engineering, as well as general interest NIST information in such important areas as calibration services, standard reference materials and data, cooperative research opportunities and grants, conferences and workshops, international standards and trade, laboratory accreditation, and policy and program changes. I hope to continue to make the *Journal* required reading for scientists, engineers, technicians, and science and technology managers who are at all concerned with measurement.

Barry N. Taylor

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

Role of Numeric Databases in Materials and Biological Sciences

Foreword

A Workshop on the "Role of Numeric Databases in Materials and Biological Sciences" was held at the annual Meeting of the American Crystallographic Association in Philadelphia, PA on 26 June 1988. The purpose of the Workshop was to provide information on scientific databases, software, and associated tools which are available in the fields of materials science and molecular biology. The results of the Workshop are being published as a series of scientific papers in this issue of the Journal of Research of the National Institute of Standards and Technology.

The Workshop was divided into two sessions. The morning session focused on the materials science databases. Applications of databases with information on electron diffraction, mass spectroscopy, thermodynamics, crystallography, and phase diagrams were discussed. Areas covered included materials characterization and design. The afternoon session centered on molecular biology databases containing numeric, factual, and bibliographic data. The applications included the use of crystallographic, nuclear magnetic resonance, sequence, and related data in the modeling and solution of three-dimensional structures.

From the Workshop, it became clear that numeric databases are playing a critical and dynamic role in the advancement of materials and biological sciences. A vast amount of data has been collected, standardized, evaluated, and organized into computer readable databases. To allow scientists to use efficiently the databases in research, mathematical and computer software tools have been developed that include database management systems, search procedures, display and graphics packages, statistical techniques, and complex analysis algorithms. Several different "user friendly" computer search and analysis systems are now in use: (1) International Online Systems in which a collection of related databases are integrated and accessed by unified search software; (2) Analytical Systems in which a database (e.g., Electron Diffraction) is integrated directly into the instrument that collects the data; and (3) Desktop Systems in which a database can effectively be searched using a combination of compact disc and personal computer.

A revolution is taking place that is making it possible for scientists to use numeric data and factual information in innovative and creative ways. This revolution is being driven by three principal factors. First, computer controlled instrumentation allows scientists to collect large quantities of high quality data. Second, many individual databases have reached a critical size making them indispensible in research. Finally, computational resources permit many database access options that range from simple search systems to the use of complex knowledge systems that contain collections of integrated databases along with sophisticated search and analysis tools.

John R. Rodgers
National Research Council Canada

Alan D. Mighell
National Institute of
Standards and Technology

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

David R. Lide, Jr.

We are pleased to dedicate this Special Issue of the *Journal of Research of the National Institute of Standards and Technology* to David R. Lide, Jr.

From 1969 until his retirement in 1988 to become Editor-in-Chief of the *CRC Handbook of Chemistry and Physics*, Dr. Lide was Director of the Office of Standard Reference Data. In this capacity, he was responsible for a national effort to develop reliable sources of physical, chemical, and materials properties data. Accomplishments towards this goal include the establishment of the *Journal of Physical and Chemical Reference Data* (for which he still serves as editor) which to date has published over 360 articles and supplements, the establishment of the National Standard Reference Database series which boasts more than a dozen titles of the most comprehensive and high quality sources of machine-readable data, and the establishment of a joint grants program with DOE and NSF, which signifies the first time agencies sponsoring fundamental research have made long-term commitments to evaluating and organizing scientific data. David Lide is also a leader in international data activities. As current President of CODATA, a scientific committee of the International Council of Scientific Unions, Dr. Lide is active in the coordination of database development in physical, biological, and geosciences. He also serves as Chairman of the Committee on Chemical Databases for the International Union of Pure and Applied Chemistry and as Chairman of the Task Force on Scientific Databases for the American Chemical Society. In June 1988 he received the Herman Skolnik Award of the American Chemical Society's Division of Chemical Information for outstanding contributions to the field of chemical information.

Many of the papers included in this Special Issue describe the applications of databases developed under Dr. Lide's guidance. Hence, it is appropriate that this Special Issue be dedicated to David R. Lide.

John R. Rodgers
National Research Council Canada

Alan D. Mighell
National Institute of
Standards and Technology

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

News Briefs

Developments

TURBULENCE WITHIN THE WALLS

Turbulence inside pipes makes the measurement of fluid flow difficult. The cost of inaccurate measurements to the petroleum and chemical process industries, for example, can amount to hundreds of millions of dollars annually. Probing this problem, two NIST researchers in the Center for Chemical Engineering have come up with a strategy to predict meter performance for non-ideal installation conditions. Their research could yield a more practical process to check or calibrate an installed flowmeter by determining in-situ the profile of the pipeflow entering the meter. NIST has formed an industry-government consortium to sponsor this research program on flowmeter installation effects.

Information on participating in the consortium is available from George E. Mattingly, 105 Fluid Mechanics Bldg., NIST, Gaithersburg, MD 20899; telephone: 301/975-5939.

NEW TECHNIQUE FOR MEASURING WAVEGUIDE LOSS

Scientists in NIST's Electromagnetic Technology Division, Boulder, CO, have developed a new technique for measuring propagation loss in optical channel waveguides used in optical communication, signal processing, and sensor applications. The technique is based on photothermal deflection effect which employs a laser beam to probe extremely small temperature changes resulting from the absorption of light. Other techniques used currently to measure waveguide losses risk damage to the guide, require special material preparation, or are subject to a large uncertainty due to randomly scattered light. The new technique avoids these difficulties and is applicable to a variety of waveguide materials.

For information, contact Aaron A. Sanders, Division 724.02, NIST, Boulder, CO 80303; telephone: 303/497-5341.

NIST, ARMY, AND GE TO COLLABORATE ON STANDARDS FOR REAL-TIME RADIOSCOPY

Thomas Siewert of the Fracture and Deformation Division and Leonard Mordfin of the Office of Nondestructive Evaluation (ONDE) negotiated an agreement with representatives of the Army Materials Technology Laboratory (MTL) and the General Electric Aircraft Engines Quality Technology Center for a joint effort to develop documentary standards (military and ASTM) on real-time x-ray radioscopy (RTR) for nondestructive evaluation (NDE). Real-time radioscopy and its companion technology, near-real-time digital radioscopy, constitute a powerful and rapidly emerging industrial inspection technique. Under the terms of the agreement, GE will develop drafts on the basis of their company Standards and MTL will provide funding to NIST to incorporate measurement considerations into the documentary standards.

This collaborative effort on documentary standards is expected to reinforce ONDE's project on measurement standards for RTR, which was initiated in FY 1987 with Siewert as the project leader. With help from Robert Placious, formerly of Center for Radiation Research, Siewert arranged a workshop and developed a questionnaire to identify industry's most pressing standards needs in RTR. The questionnaire generated considerable interest, including numerous invitations for Siewert to speak at NDE meetings, and leading to the collaboration. As a corollary benefit to NIST, Siewert has been granted access to MTL's and GE's RTR laboratory facilities for research purposes. GE's facilities in Cincinnati, Ohio, in particular, are among the most advanced in the country.

PATTERN-RECOGNITION TECHNIQUES APPLIED TO XRF ANALYSIS

Scientists from the NIST Center for Radiation Research and from NASA are experimenting with a new approach for the interpretation of x-ray fluorescence (XRF) spectra. XRF analysis is used widely in science and industry for the nondestructive determination of the chemical composition of a sample. In many field and quality-control applications, quantitative information on the composition is not needed. What is needed is the ability to monitor changes in the composition among samples or to select and classify samples with similar compositions. The need for XRF classification and collection of diverse geological samples by a rover vehicle on NASA's proposed Mars Sample Return Mission stimulated this investigation.

Traditionally, quantitative analysis of XRF spectra involved peak fitting (with background subtraction) to determine the areas of the characteristic x-ray peaks of interest and a time-consuming numerical conversion of these areas to chemical composition, taking into account alterations of the peak intensities by self-absorption and secondary emission in the sample. This method required knowledge of the excitation source spectrum and the peak intensities from a suite of known samples. The pattern-recognition method simply correlates selected peak regions of the raw XRF spectra with those from standard samples obtained with the same excitation source (information on the source spectrum is not needed), and this takes only a few seconds on a personal computer. A proof of principle has been demonstrated for geological and alloy samples using a field-quality system comprised of a small battery-operated x-ray generator and an energy-dispersive spectrometer.

FIPS FOR POSIX APPROVED

The Secretary of Commerce approved the Federal Information Processing Standard (FIPS) for POSIX (Portable Operating System Interface for Computer Environments). To be issued as FIPS 151, the standard has been adopted on an interim basis to enable the Federal Government to use the POSIX specification in procurements and in developing systems for applications portability.

FIPS 151 adopts draft 12 of the Institute of Electrical and Electronics Engineers (IEEE) Standard for POSIX. A FIPS adopting final voluntary standard specifications for POSIX will be proposed when those specifications are completed.

As currently defined, POSIX is the crucial first step in providing a vendor independent interface specification between an application program and an operating system. However, the current definition must be extended to provide interface specifications for full operating system functionality.

In addition to a fully extended POSIX that supports source code portability across many different machines and operating systems, there is a need for an architectural approach to applications portability. National Computer and Telecommunications Laboratory (NCTL)—formerly ICST—is working with industry and users to produce the needed specifications for both the extended POSIX and an Applications Portability Profile (APP).

The APP will be a group of standard elements including database management, data interchange, network services, user interfaces, and programming languages. Workshops were held in September, October, and November 1988 and others are scheduled for January and May 1989 to discuss the APP and the POSIX standard.

AGREEMENT WITH NCC FOR VALIDATION OF FIPS COBOL AND FORTRAN

The NCTL and the National Computer Centre (NCC) of the United Kingdom signed an agreement to recognize COBOL and FORTRAN test reports, and validation certificates issued by each other. As the basis for mutual recognition of test reports and validation certificates, NCTL and NCC agree to use the same test method, to follow equivalent validation procedures and to adopt equivalent certification criteria. Under the agreement, NCTL can also recognize testing done by other European Economic Community or the European Free Trade Association test centers that are sublicensed by NCC.

The test method used for mutual recognition of FORTRAN validations is NCTL's FORTRAN Compiler Validation System (FCVS) for testing FIPS 69-1 (ANS X3.9-1978) FORTRAN. The test method used for mutual recognition of COBOL validations is the COBOL Compiler Validation System (CCVS) for testing FIPS 21-2 (ANS X23-1985) COBOL. The CCVS was initially developed by the U.S. Government and later updated for the 1985 COBOL standard (FIPS 21-2) by NCC under a joint project with the U.S. Government, United Kingdom, France, and West Germany.

LARGE-SCALE CRACK-ARREST TESTS ON REACTOR STEEL COMPLETED

With the final test performed on Sept. 22, 1988, a 5-year program sponsored by the Nuclear Regulatory Commission involving the fracture of large plates of reactor grade steel was completed on schedule. The NIST work provided data required for the evaluation of one of the main concerns in nuclear plant operation: protecting the main pressure vessel against brittle fracture. Significant accomplishments by NIST: (1) extended the existing limits on crack-arrest databases to regions of current engineering interest; (2) clearly established that brittle crack-arrest does occur prior to ductile crack extension; (3) improved elastic and viscoplastic fracture mechanics models; and (4) developed improved dynamic fracture methods. Scientific advances were made in the area of measurement of crack inertia and plastic zone growth kinetics. A 2-year analytical program is planned to explore certain research issues that were raised by the completed testing program.

INDUSTRIAL WORKSHOP ON INTELLIGENT PROCESSING OF MATERIALS

A 2-day workshop was held at Gaithersburg, MD for more than 50 industrial representatives in order to assess the priorities for research in the important emerging technology of intelligent processing of materials. Another aim of the workshop was to initiate discussion on possible cooperative NIST/industrial projects. The workshop was sponsored by the Institute for Materials Science and Engineering and the Office of Nondestructive Evaluation with participation by the Ceramics, Metallurgy, Fracture and Deformation, and Polymers Divisions. The industrial participants, representing both materials producers and users, assessed the needs for intelligent processing technology for a range of advanced materials including ceramics, metal alloys and polymers. The proceedings of the workshop will provide valuable planning information for joint NIST industrial research on intelligent processing of materials.

EXPANSION OF THE NIST-DOD FLOWMETER TESTING PROGRAM

The national standards laboratories in Italy and the United Kingdom have joined the round-robin flowmeter testing program being conducted by the NIST Chemical Process Metrology Division and sponsored by the Department of Defense (DOD). Tests have shown that fluid flow measurements in two European laboratories—the Instituto di Metrologia in Italy and the National Engineering Laboratory in Scotland—show such good traceability links to NIST results that overseas DOD labs, such as the U.S. Navy in the Mediterranean and the U.S. Air Force in the United Kingdom, can use these laboratories as sources for calibrations. The ultimate goal of the program is to quantify the traceability of all DOD flow measurements to NIST, which will provide DOD with more widely distributed and convenient sources of flow measurement services.

ACCESS CONTROL RESEARCH

Researchers are designing a prototype system to protect the confidentiality and integrity of information in local area networks (LANs) consisting of workstations and host computers. To gain access to network resources, users sign on automatically using smart tokens (smart cards, keys, or modules) inserted into a card reader/writer attached to the workstations and host computers. A passive token stores a password or cryptographic information to verify the user's identity. Smart cards with computation capabilities can be used in both user-to-host and host-to-user authentication processes.

Smart Card Technology: New Methods for Computer Access Control, (NIST Special Publication 500-157), describes the basic components of a smart card and provides background information on the underlying integrated circuit technologies. The capabilities of a smart card are discussed, especially its applicability for computer security. The report describes research being conducted on smart card access control techniques; other major U.S. and international groups involved in the development of standards for smart cards and related devices are outlined in the appendix.

APPROVAL OF FIPS PUB 152, STANDARD GENERALIZED MARKUP LANGUAGE

On Sept. 26, 1988, the Secretary of Commerce approved the standard for Generalized Markup Language (SGML) to be published as Federal Information Processing Standard Publication (FIPS PUB) 152. Effective March 31, 1989, the new standard adopts the International Standards Organization SGML (ISO 8879-1986) which specifies a language for describing documents to be used in office document processing, interchange between authors and between authors and publishers, and publishing. The language provides a coherent and unambiguous syntax for describing the elements within a document.

The new SGML standard provides a common markup language for a variety of document types and uses, permitting the portability of unformatted textual data among different installations and processing systems and promoting interchange of documents between systems of different manufacturers. The standard is appropriate for documents which are processed by any text processing system.

CCAM DEVELOPS INSULATION ECONOMICS PROGRAM

At the request of the U.S. Department of Energy, Steven Petersen of CCAM's Mathematical Analysis Division has developed and field tested the Zip-Code Insulation Program (ZIP), a microcomputer program for determining economic levels of insulation in new and existing houses. ZIP determines economic insulation levels for attics, walls, floors, basements, crawlspaces, and slab edges, based on local climate conditions, energy costs, and insulation costs, all keyed to the first three digits of the user's Zip Code. This program and supporting data files will be used by utilities, insulation manufacturers and vendors, energy specialists, and homeowners. ZIP will serve as the primary reference for DOE's insulation guidelines for homeowners.

CBT AND CCE HOST INDUSTRY WORKSHOP ON PROPERTY DATA FOR OZONE-SAFE REFRIGERANTS

Approximately 35 experts from industry, government, and universities met at NIST on Sept. 22, 1988, to identify needs for thermodynamic and transport property data for replacements for those refrigerants that damage the ozone layer in the upper atmosphere. The working fluids for much of small and large-scale heating and cooling equipment and foam blowing of thermal insulation are currently chlorofluorinated hydrocarbons (CFCs). Of the fully halogenated CFCs the most common are trichlorofluoromethane (R11) and dichlorodifluoromethane (R12). Leakage of R11 and R12 beyond the troposphere is destroying the ozone layer. The environmental ramifications of the continued use of R11 and R12 are so pressing that the United States signed the Montreal Protocol in December 1987 with 31 of the world's major producing and consuming countries. The agreement provides for a near-term freeze on the manufacturing of R11 and R12 followed by scheduled 50 percent reductions in their manufacture. It is unlikely that industry will be able to replace R11 and R12 in existing equipment without modifications to this equipment. The design of high-quality new equipment will depend critically on the thermophysical properties of the replacement refrigerant materials.

MODIFICATIONS IMPROVE NIST GAS FLOW FACILITY

Sellers and buyers of natural gas and other gas products will have greater confidence in their transactions as the result of improvements to NIST's gas flow measurement facility at its Boulder, CO laboratories. The facility, first put into operation in 1979, measures the performance of flowmeters that assure the accuracy of transactions between sellers and buyers. As a result of the improvements, variability in gas temperature has been decreased by a factor of five and the precision of performance data on flowmeters has improved by a factor of two. This increased precision has made the facility much more useful and capable of performing a wide variety of research, says a recent report on the improvements. Major changes were made to the gas flow loop, calibration lines, and the regulator for the gas supply to the pneumatic controllers. In addition, a new minicomputer replaced three small computers to improve data-analysis capability, and an additional cooling line was installed to supply extra liquid nitrogen to the main heat exchanger, improving temperature control.

A copy of the paper outlining the improvements is available from Fred McGehan, Public Affairs Office, NIST, Boulder, CO, 80303. More information on the facility is available from Susan E. McFaddin, Chemical Engineering Science Division, NIST, Boulder, CO 80303.

ECONOMICS, EFFICIENCY OF INSULATION WITHOUT CFC

Two of the most efficient insulating products used in new building construction today are polyurethane and extruded polystyrene rigid foam insulation; both contain chlorofluorocarbons—CFCs. During the foam manufacturing process, CFCs are used to form gas cells or bubbles in the foam making it an excellent insulator. However, there is evidence that CFCs can break down the Earth's ozone layer, and their manufacture and use may be curtailed. In a study for the Department of Energy, NIST researchers looked at the cost-effectiveness and potential energy consequences of using expanded polystyrene (EPS) and fiberglass—neither contains CFCs. They found that both insulation materials typically cost less than most CFC-containing foams. But since they contain air bubbles instead of gas bubbles, they do not insulate as efficiently. And, because more of the material is

needed to achieve the same thermal performance, there may be an increase in cost if walls or roof areas must be expanded to accommodate the thicker insulation.

A report, Technical and Economic Analysis of CFC-Blown Insulations and Substitutes for Residential and Commercial Construction, is available from the National Technical Information Service, Springfield, VA 22161, for $14.95 prepaid. Order by PB #88-243399.

QUANTUM EFFECTS DOMINATE 1989 GRANTS

NIST announced recently that it will award two 1989 Precision Measurement Grants to Randall G. Hulet of Rice University and to Edward Hinds and Malcolm Boshier of Yale University. The grants, for $30,000 each for fiscal year 1989, will support Hulet's study of atomic collision processes at extremely low temperatures, and Hinds and Boshier's precision measurement of energy levels in hydrogen and helium atoms. Both experiments are designed to probe fine details of quantum theory. The NIST Precision Measurement Grants program was started in 1970. The awards are made annually to scientists in academic institutions for work in the fields of precision measurement and the study of fundamental constants of nature. Each grant is awarded for 1 year and may be renewed for up to 2 additional years at the discretion of NIST.

PROPOSALS WANTED FOR 1990 GRANTS

NIST is seeking project proposals for two research grants for fiscal year 1990 in the field of precision measurement and fundamental constants. The Precision Measurement Grants are for $30,000 for 1 year, and may be renewed by NIST for up to 2 additional years. Prospective candidates must submit summaries of their proposed projects to NIST by Feb. 1, 1989, to be considered for the current grants, which will run from October 1989 through September 1990. NIST Precision Measurement Grants are awarded each year to scientists in academic institutions for work in determining values for fundamental constants, investigating related physical phenomena, or developing new, fundamental measurement methods. The grants were instituted in 1970 to augment NIST research programs in physical constants and fundamental measurements, and to encourage research in these fields at colleges and universities. To date, 44 grants have been awarded.

For further information, contact Dr. Barry N. Taylor, NIST Precision Measurement Grants Committee, B258 Metrology Bldg., NIST, Gaithersburg, MD 20899; telephone 301/975-4220.

Standard Reference Materials

STANDARD REFERENCE MATERIALS 3191-3195 AQUEOUS ELECTROLYTIC CONDUCTANCE STANDARDS

The Office of Standard Reference Materials announces the availability of a series of SRMs certified for aqueous electrolytic conductance. The series, SRMs 3191-3195, has nominal conductivities that range from 100 to 100,000 micro-siemens per centimeter at 25 °C. These SRMs are intended primarily for use as a control and in the calibration and standardization of conductivity cells and meters used in water purity determinations. Indications of water purity can be determined in this way since ionic impurities of only a few parts per million can be detected readily by measuring the conductivity of the water. The solutions are prepared by dissolving high-purity potassium chloride in deionized water in equilibrium with atmospheric carbon dioxide.

Certification of these conductance standards was performed in the Inorganic Analytical Research Division of the Center for Analytical Chemistry.

ELLIPSOMETRIC SRM A VIRTUAL SELL-OUT

Industrial firms have ordered nearly all of the first available 57 units of Standard Reference Material (SRM) 2530 for ellipsometrically derived thickness and refractive index of a silicon dioxide layer on silicon. SRM 2530 was developed by CEEE's Semiconductor Electronics Division in response to the needs of the semiconductor (and other) industries to evaluate the accuracy of ellipsometers, but it may also be used as aid in the calibration of various other optical and mechanical thickness monitoring instruments.

The SRM is available separately for three oxide thicknesses: 50 nm (2530-1), 100 nm (2530-2), and 200 nm (2530-3). All three thicknesses have proven to be about equally popular. The division has fabricated about 150 of the standards and in response to the continuing demand is completing the characterization of the remainder for certification using

the highly accurate ellipsometer designed and built in the division.

Each SRM consists of a 76-mm diameter silicon wafer on which a uniform silicon dioxide layer was grown, patterned, and partially covered with chromium. Certified values are provided for the derived values of thickness and refractive index of the silicon dioxide layer, as calculated by means of a model which postulates the existence of a thin silicon-rich oxide interlayer as well as the silicon dioxide layer. The SRM package includes a copy of Preparation and Certification of SRM 2530, Ellipsometric Parameters Delta and Psi and Derived Thickness and Refractive Index of a Silicon Dioxide Layer on Silicon (NIST Special Publication 260-109) and a FORTRAN program to aid the end-user in computing the ellipsometric parameters and the thickness of the oxide.

STANDARD REFERENCE MATERIALS 3171-3176—MULTIELEMENT SOLUTION STANDARDS

The Office of Standard Reference Materials announces the availability of high-purity Standard Reference Materials (SRMs 3171-3176) that should provide improved accuracy for multielement techniques or any technique that requires aqueous solutions for calibration and will enable better trace element analysis. These SRMs will complement the existing single-element SRMs (3101-3169) and are prepared from high-purity metals, salts, and reagents.

The first two multielement solutions, SRM 3171, Multielement Mix A Standard Solution, and SRM 3172, Multielement Mix B Standard Solution, were available in November, 1988 and each consists of 10 elements in a 5-percent nitric acid solution. The remaining four mixes were available in December 1988. The elements in the solutions are:

SRM 3171 – Al, Ve, Cd, Cr, Fe, Mg, Mn, Ni, K, Na
SRM 3172 – As, Be, Ca, Co, Cu, Pb, Se, Ag, Sr, Zn
SRM 3173 – Bi, B, Ce, Hg, P, Si, Tl, Th, V
SRM 3174 – Al, Be, Cd, Hf, Fe, Zr, Au, Ti, Pb, B
SRM 3175 – As, Be, La, Mo, P, Se, Te, Sn, Y
SRM 3176 – Sb, Be, B, Dy, Eu, Gd, Nd, Ru, Ti

Certification of these standards was performed in the Inorganic Analytical Research Division of the Center for Analytical Chemistry.

STANDARD REFERENCE MATERIAL 4339 RADIUM-228

Standard Reference Material (SRM) 4339 has been developed for accurate calibration monitoring of instruments for radium-228, following 5 years of extensive sample preparations and different calibration methods. The monitoring of this natural environmental radioactivity, the most significant radium contaminant in some areas, is required by the Clean Water Act.

The radium-228 in this SRM is from 25-year-old purified thorium oxide and contains no detectable radionuclides other than radium-228, $<.01$ percent radium-226 and their expected progeny. Activity measurements are difficult for the 5.75-year half-life radium-228 because of the low beta-particle energies and the high-energy emissions from daughter 6.13-h ^{228}Ac and subsequent progeny. Rapid and simple chemical separations are complicated by the presence of 3.6-day ^{224}Ra later in the chain.

This new Standard Reference Material consists of quantitative samples of calibrated radium-228 and 5-mL of carrier solution in a flame-sealed glass ampoule, with an overall uncertainty of less than 3 percent.

Certification of SRM 4339 was performed in the Ionizing Radiation Division, Center for Radiation Physics.

Standard Reference Data

CHEMICAL THERMODYNAMICS DATABASE AVAILABLE ONLINE

The computerized version of the NBS Tables of Chemical Thermodynamic Properties now is available worldwide to subscribers of STN International (Scientific and Technical Network). The numerical data in the new file provides researchers in chemistry, physics, and manufacturing, and environmentalists with rapid access to evaluated data on more than 15,000 inorganic substances. A complete description of the chemical system is given with values for six thermodynamic properties: enthalpy of formation, Gibbs energy of formation, heat capacity and entropy at 298.15 kelvin, enthalpy differences between 298.15 kelvin and 0 kelvin, and the enthalpy of formation at 0 kelvin. This is the second NIST database to be offered online through STN, a retrieval service jointly offered by the American Chemical Society, the

Fachinformationszentrum Karlsruhe (FIZ Karlsruhe, West Germany), and the Japan Information Center of Science and Technology.

For information on the availability of the database through STN, contact the Office of Standard Reference Data, A323 Physics Bldg., NIST, Gaithersburg, MD 20899, Telephone: 301/975-2208.

PROPOSALS REQUESTED FOR 1989 GRANTS PROGRAM ON STANDARD REFERENCE DATA

Project proposals are requested by the National Institute of Standards and Technology (NIST)—formerly the National Bureau of Standards—to compile and evaluate scientific reference data for scientists and engineers to use in research, development, and the design of industrial processes.

Established in 1980, the grants program is administered by the NIST Office of Standard Reference Data (OSRD), which will award approximately $400,000 provided by the National Science Foundation and NIST. Typically, project proposals are funded at $20,000-40,000 for 1- to 2-years. The closing date for receipt of the proposals is Feb. 28, 1989.

Proposals will be considered which are concerned with the physical, chemical, or materials properties of well-characterized substances or systems. Projects may cover several properties of a single substance or a single property of a coherent group of substances.

The work proposed should include the collection of data from the scientific literature and the critical evaluation of those data. Each project must lead to a publishable compilation, critical review, or computer database containing recommended values with stated uncertainties for the properties in question.

Because the program emphasizes the critical evaluation of published data, proposals which involve only the compilation of data without the exercise of scientific judgment will not be considered. Projects should not involve new experimental measurements.

Examples of appropriate subjects for the data compilations include: thermochemical data and properties of aqueous solutions; thermophysical properties of fluids; mechanical properties of metals, ceramics, polymers and composites; chemical kinetic data; atomic structure and collision data; molecular structure and dynamics data; data on surface characterization; corrosion-related data; semiconductor or superconductor properties; and spectral data for chemical analysis.

The projects are expected to have a well-defined goal and be sufficiently limited in scope to produce useful results in 1 or 2 years.

The 1989 grants program is open to researchers in any U.S. organization, academic or non-academic, non-profit or commercial. Judgments on the proposals will be made on the basis of the importance of the data to be collected, the feasibility of the project, the qualifications of the investigator, and the possibility of other support for the project. Those interested in applying for a grant under this program should contact the Office of Standard Reference Data, A323 Physics Bldg., National Institute of Standards and Technology, Gaithersburg, MD 20899, telephone: 301/975-2200.

Calendar

July 26–August 1, 1989

16th INTERNATIONAL CONFERENCE ON THE PHYSICS OF ELECTRONIC AND ATOMIC COLLISIONS

Location: Grand Hyatt Hotel
New York, NY

This biennial international conference seeks to promote the growth of scientific knowledge and its effective exchange among investigators of all nations in the field of electronic and atomic collisions and related areas of atomic and molecular physics. The conference deals with two-body interactions between ions, atoms, molecules, electrons, positrons, and photons. The conference is cosponsored by the International Union of Pure and Applied Physics, the National Institute of Standards and Technology, the American Physical Society, the U.S. Department of Energy, the U.S. Air Force Office of Scientific Research, the U.S. Office of Naval Research, and the National Science Foundation.

For further information about the conference, contact Thomas Lucatorto, A251 Physics Building, NIST, Gaithersburg, MD 20899, 301/975-3734.

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

September 19–21, 1989

12th INTERNATIONAL SYMPOSIUM ON POLYNUCLEAR AROMATIC HYDROCARBONS

Location: National Institute of Standards and Technology Gaithersburg, MD

The first eleven symposia on the chemistry and biochemistry of polynuclear aromatic hydrocarbons (PAH) were held on an annual basis and proved to be the major focal point for multidisciplinary research on this important class of chemical species. The meetings are now held biannually and continue to encourage open discussions between scientists representing government, academic institutions, industry, and research facilities investigating the chemical properties and biological effects of these compounds. The 1989 meeting will include presentations on parent hydrocarbon PAH as well as het-eroatomic species, and PAH derivatives including amino, nitro and halogen substituted compounds. Topics to be discussed include adducts, bioactivity, carcinogenesis, mutagenicity, cell transformation, detoxification, epidemiology, pollution modeling, occupational exposure, organic synthesis, and environmental studies. Sponsored by the National Institute of Standards and Technology, the National Institutes of Health, and Battelle Memorial Institute.

Contact: Willie E. May, A113 Chemistry Building, NIST, Gaithersburg, MD 20899, 301/975-3108.

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

The Importance of Numeric Databases to Materials Science

Volume 94 Number 1 January–February 1989

Richard A. Matula

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

Scientific numeric databases are important research tools for materials scientists. In distinction to bibliographic databases, these numeric databases are useful primarily to provide direct, immediate access to data, often evaluated data. Examples showing the application of crystallographic databases are given, including determining candidate materials for certain applications. Thermochemical data useful for optimizing optical fiber processing are discussed showing the importance of high-quality data. In addition, these databases are an important tool that can be utilized in the graduate education of the next generation of materials scientists.

Key words: CRYSTDAT; industrial use of numeric databases; lattice matching; materials science; numeric databases; online searching; prediction of ferroelectricity; superconductivity.

Introduction

In this introduction, I would like to briefly point out some aspects of the importance of numerical databases to materials sciences, specifically the importance of good data and the quick dissemination of numeric data to a wide audience. All the examples that are mentioned come from the work at AT&T Bell Laboratories.

If one were to claim that numeric databases were the most important development in materials science, that would clearly be an overstatement. On the other hand, if one were to claim that they are of no value to materials science, that would be understating their importance. The importance lies in between and can be expected to increase as knowledge of them becomes more widely known. Many of these databases, specifically in the crystallographic area, have developed from print products that existed for decades and putting them into numeric databases is a shift in content and in means of access. But it's more than that, much more than that. As with many new technologies, there are uses that are unanticipated when a new technology first appears. The usual advantages of having data in numeric form include the following: they're easy to update, there's a potentially faster delivery of information to the user, and there's a potentially large audience of scientists and engineers that can be reached by having the data in the electronic form, because these people can access this data by their personal computer or microcomputer in the laboratory, the office, or even at home.

I want to point to several examples. First, I'll discuss one of the main processes for making optical fibers, which I'll spend the most time on, and which will illustrate the need for good data. Other examples I will mention are the prediction of ferroelectricity, lattice matching, and high-T_c superconductivity. I'll close with a lesson from history.

Modified Chemical Vapor Deposition (MCVD) for Optical Fibers

This first example[1] will show the crucial importance of good data to industry. This concerns modeling the process involved in the manufacture of optical fibers.

Figure 1 shows a part of a highly purified piece of glass rod that would be used to make an optical fiber. One can observe hardly any scattering coming from the laser beam (which goes from the upper left to the lower right) going through the glass rod. This is due to both the high purity and the absence of scattering centers. There is a high index of refraction core in the center of the rod. This core is formed by doping silica glass with germanium and phosphorus. When an optical fiber is drawn from the rod, this core serves to guide the laser beam as it travels through the fiber. The rod or preform shown in figure 1 was made by a process known as the MCVD process developed by J. B. MacChesney and others at AT&T Bell Laboratories in the mid-1970s [3]. Two other processes have been developed to produce similar fibers; they are the OVD (outside vapor deposition) process, otherwise known as the Corning process, and the VAD (vapor phase axial deposition) process, which is mainly used in Japan.

Figure 2 is a schematic representation of MCVD. This shows a silica tube being rotated in a lathe. Reactant gases enter on the left, pass through, and are exhausted. An oxyhydrogen torch heats the glass tube together with the reactant gases inside, and the torch travels slowly the length of the tube in the direction indicated. In the heated zone (about 1650 K), silicon tetrachloride and other halides react with oxygen to form oxide particles, and slightly downstream these particles deposit on the inside wall of the tube [4] and are vitrified as the torch passes over them. The tube will later be collapsed and then drawn into a hair-thin optical fiber many kilometers long.

Figure 3 shows the principal species involved in the reactions in the heated zone, namely SiO_2, GeO_2, etc., together with 32 gaseous and liquid species. Thermochemical data for these species needed for modeling the process came from several sources [5,6,7,8]. The existence and concentration of these species were imperfectly known in the minds of researchers for some time. Empirical work, together with modeling, led to a better understanding of the process including better values of equilibrium constants [9]. What is important is the formation of GeO_2 from $GeCl_4$. Both species are present at equilibrium and small changes in deposition conditions can significantly change the concentration of each.

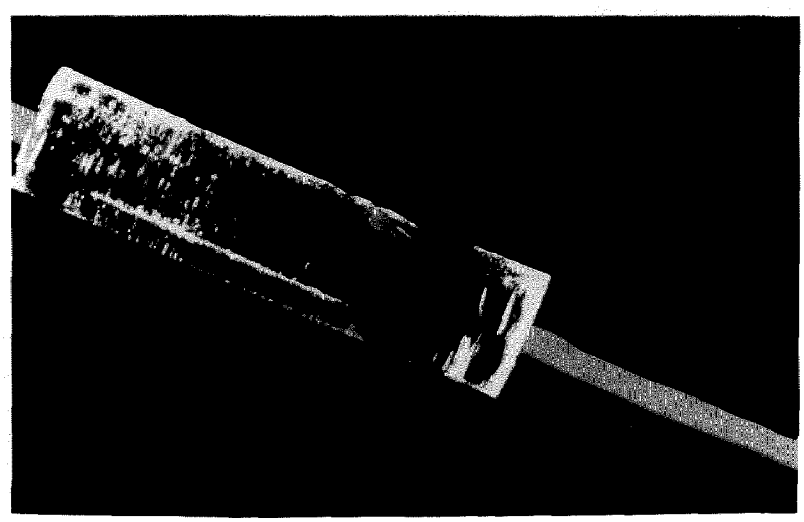


Figure 1. Scattering not observed in highly purified piece of glass rod.

[1] For this example, I want to thank K. B. McAfee, Jr., who recently retired from AT&T Bell Laboratories. He presented this example in a talk, which is unpublished, at the Tenth CODATA Conference in Ottawa, July 1986. Credit for this work goes to K. B. McAfee, Jr., R. A. Laudise, R. S. Hozack, D. M. Gay, and K. L. Walker all of AT&T Bell Laboratories [1,2].

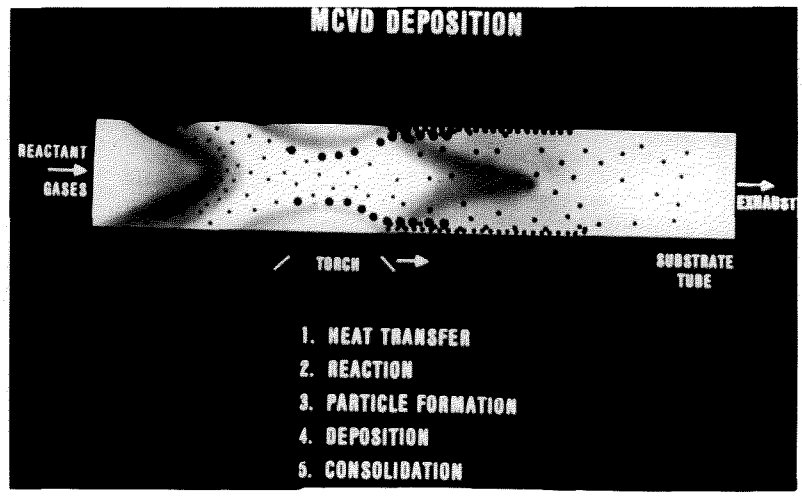


Figure 2. Essence of MCVD process.

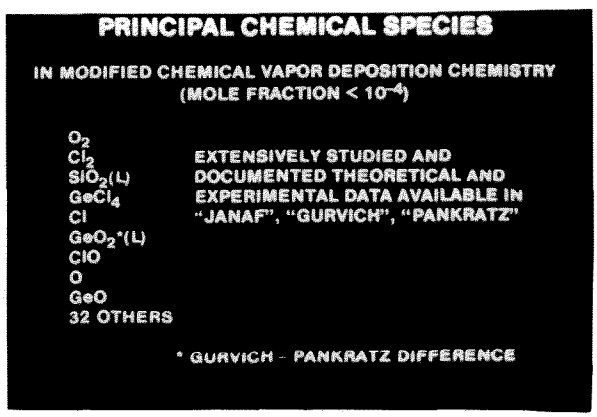


Figure 3. Principal chemical species in MCVD chemistry.

Figure 4 shows the mole fraction of the various reactants and products as a function of temperature. The bump in the GeO_2 curve was predicted by the modeling of the process and was later verified experimentally. Figure 5 shows the mole fraction of GeO_2 in the liquid phase as a function of the Ge/Si ratio in the feed gas. The two dotted lines and the solid line in between come from the model calculations for various temperatures. The dots and error bars come from experimental data. The heat of formation that was chosen was such as to get the best agreement between the modeling and the experimental data. However, another reputable researcher has heat of formation data, leading to the upper curve, that differs by about 4 kilocalories compared to the heat of formation data that fits the modeling and experimental data very well. This shows the need for heat of formation data to better than 1 kilocalorie instead of 4 kilocalories. A more extensive discussion of this difference appears in the literature [10].

This example illustrates that better values of data, when placed in the proper theoretical and experimental context, can lead to a more thorough understanding and an optimization of an industrial process. This optimization has further implications when it is realized that the AT&T plant in Atlanta is capable of producing a large mileage of fiber per year, and hence is of significant commercial importance.

Predictions of Ferroelectricity

Numerical databases can be used in novel ways. The previous example would exemplify the usual retrieval of numerical data for a certain material. Ten years ago, when I was at CINDAS (Center for Information and Numerical Data Analysis and Synthesis, Purdue University), most of the requests for retrieval of data were just of this kind. For example, one would ask for resistivity of palladium over a certain temperature range. At that time we thought of the potentiality of reversing that type of search. In the reverse type of search one would ask for a list of materials fitting a certain set of conditions. For example, in the design of a spacecraft, what materials would have both a flat thermal linear expansion over a certain temperature range and a thermal emissivity between two given values? That illustrates the use of numeric databases for a

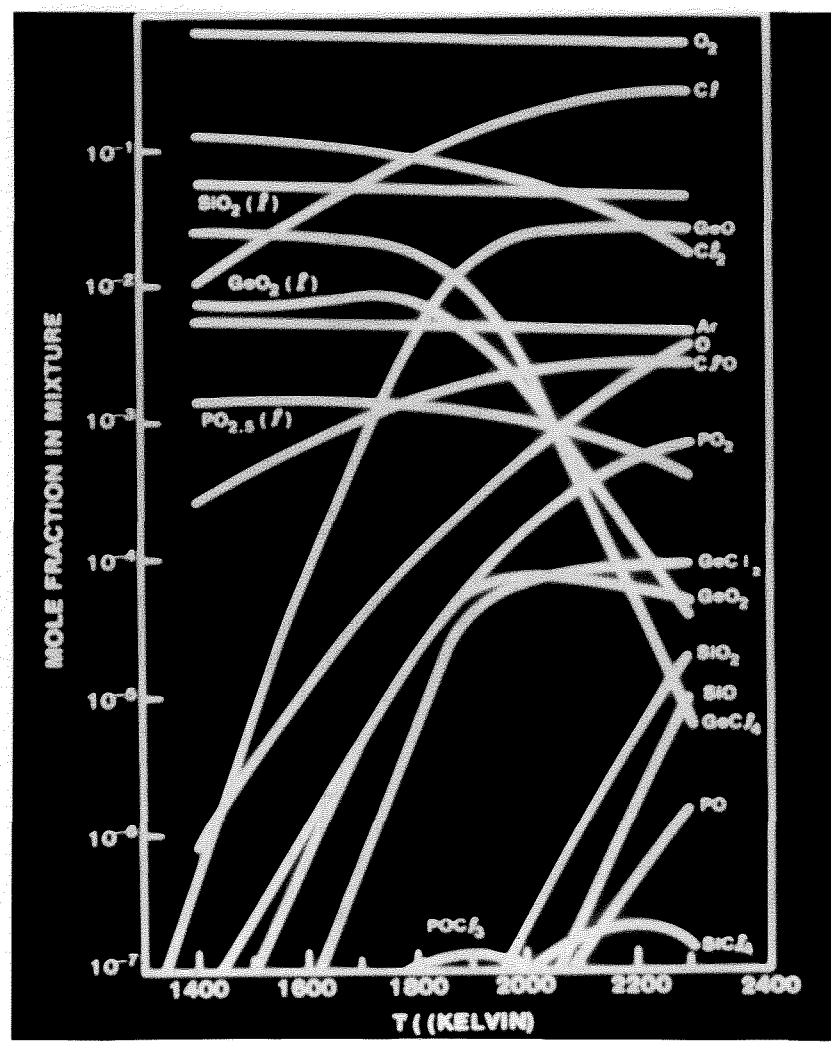


Figure 4. Mole fraction of chemical species in MCVD process.

new type of question which we are now able to answer, and S. C. Abrahams of AT&T Bell Labs has discussed how he utilized the Inorganic Crystal Structure Database (ICSD) mounted in Karlsruhe, West Germany, to predict ferroelectricity in materials from point group 6*mm* using a structure-based approach [11]. Using the database, together with the criterion of polar space groups, the number of structures that needed to be examined was considerably reduced; additional structural criteria were then used to make the final predictions. He points out that experimental verification of each prediction is needed. The approach of using a crystallographic database for this type of query is important and advances such forward-looking research.

Lattice Matching

Another example of this reverse approach is that of lattice matching. If one has, for example, gallium arsenide and wants to make sure that a material for lattice matching has not been overlooked, these crystallographic databases can be searched to determine what materials have a certain lattice parameter very close to that of GaAs.

High-T_c Superconductivity

Theo Siegrist discusses applications of the crystallographic database, CRYSTDAT, to high-T_c superconductivity research [12] with the result of a

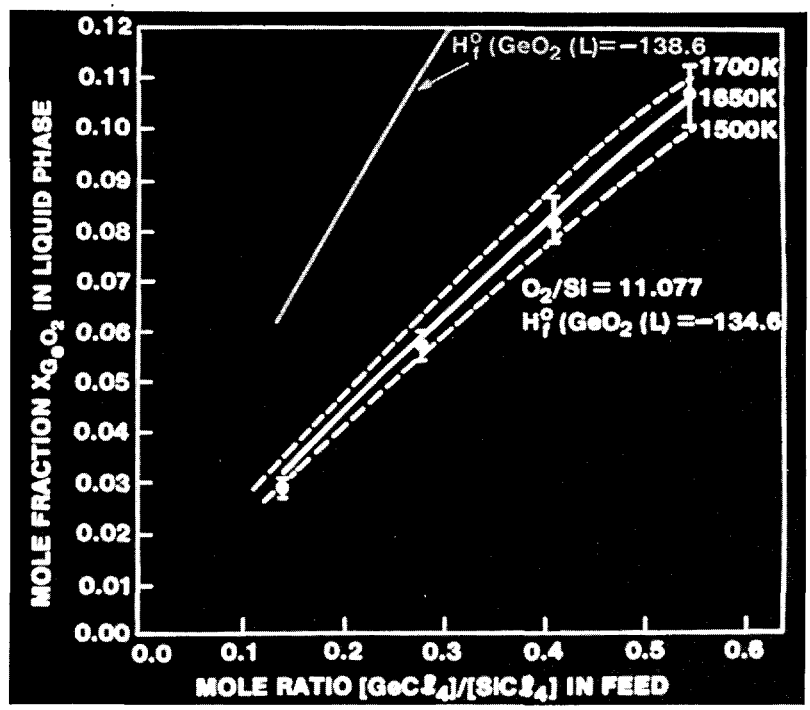


Figure 5. Difference of 4 kilocalories in heat of formation of germanium dioxide important for modeling MCVD process.

substantial time saving in determining a crystal structure. He and his colleagues were able to determine the crystal structure quickly, in about a day, compared to other groups having to spend a week or so. Clearly, these researchers were more productive because their time was used more efficiently.

It should be pointed out that information on crystallographic databases appears in *Crystallographic Databases* that has been recently published [13]. It contains a discussion of many other crystallographic databases in addition to the two referred to in this talk.

Lesson from History

There's a lesson from history as we look back at different types of databases, mainly bibliographic, that may be applicable to the work in numeric databases. The publisher of *Chemical Abstracts* has for a long time had an educational policy that encourages the use of their online database by students. The students get training while in graduate school, and when they come to industry, they expect to have this tool available. In the legal area, a similar practice was done for lawyers by Mead Data and their Lexis system of databases. In order to have the next generation of materials scientists use these databases to the maximum degree possible, how would the following question be answered: Is everything possible being done to make sure that the graduate schools are using these numerical databases in the training of materials scientists?

For the two examples mentioned using the crystallographic databases ICSD and CRYSTDAT, a personal password was obtained for each of the two researchers so that they could access the database at their convenience. What other actions, including advertising and publicity, can information professionals do to increase knowledge of these databases, and other numeric databases, in order to encourage increased usage and hence aid the technical staff to become more effective?

For present users, what other numerical databases are useful? What new combinations of data should be put together in existing databases? With the technology of split screens and microcomputers, is it possible to include additional computational power in the databases? I know John Rodgers (613-993-3294) of the National Research Council Canada, and Alan Mighell of NIST (301-975-6255) who are the conference organizers, are

very interested in this question, and would like to know your thoughts. In general, other database producers would very much like to have comments from users, to determine how their databases should develop.

Lastly, what policies are in place, or if not, should be in place, to ensure that the critical work in database production and data evaluation is passed on to the next generation? What needs to be done so the torch, so to speak, does not go out because of retirement, death, or other untimely changes? It is important to ensure that the work that's being done in and on these databases does not go for naught and can continue.

Acknowledgments

I would like to thank the following individuals for valuable discussions: S. C. Abrahams, K. B. McAfee, Jr., T. Siegrist, and especially J. B. MacChesney. The discussion presented here, however, is the responsibility of the author.

References

[1] McAfee, K. B., Jr., Laudise, R. A., and Hozack, R. S., Equilibria Concentrations in the Oxidation of $SiCl_4$ and $GeCl_4$ for Optical Fibers, J. Lightwave Tech. **LT-1**, 555 (1983).

[2] McAfee, K. B., Jr., Gay, D. M., Walker, K. L., and Hozack, R. S., Thermodynamic Stability and Incorporation of Phosphorus Into Germanium-Doped Silicon Glass, J. Amer. Ceram. Soc. **68**, 359 (1985).

[3] Nagel, S. R., MacChesney, J. B., and Walker, K. L., Modified Chemical Vapor Deposition, Chapter 1 (pages 1-64) in *Optical Fiber Communications—Volume 1: Fiber Fabrication*, T. Li (Editor), Academic Press, New York (1985).

[4] Simpkins, P. G., Greenberg-Kosinski, S., and MacChesney, J. B., Thermophoresis: The Mass Transfer Mechanism in Modified Chemical Vapor Deposition, J. Appl. Phys. **50**, 5676 (1979).

[5] Stull, D. R., et al., JANAF Thermochemical Tables, NSRDS-NBS37, 2nd Ed, U.S. Dept. of Commerce (June 1971). [The JANAF thermochemical database is now available through the STN system (the Scientific and Technical Information Network, with the North American Node in Columbus, Ohio.)]

[6] Chase, M. W., JANAF Thermochemical Tables Magnetic Tape Version, Supplement #55, Dow Chemical Co, Midland, MI (1982).

[7] Gurvich, L. V., et al., Thermodin, Svoistva Individual'nykh Veshchestv, Volume II, Academy of Sciences of U.S.S.R. (1979).

[8] Pankratz, L. B., Thermodynamic Properties of Elements and Oxides, U.S. Bureau of Mines Bulletin 672 (1982).

[9] Wood, D. L., Walker, K. L., MacChesney, J. B., Simpson, J. R., and Csencsits, R., Germanium Chemistry in the MCVD Process for Optical Fiber Fabrication, J. Lightwave Tech. **LT-5**, 277 (1987).

[10] McAfee, K. B., Jr., Walker, K. L., Laudise, R. A., and Hozack, R. S., Dependence of Equilibria in the Modified Chemical Vapor Deposition Process on $SiCl_4$, $GeCl_4$, and O_2, J. Amer. Ceramic Soc. **64**, 420 (1984).

[11] Abrahams, S. C., Structurally-Based Prediction of Ferroelectric-Paraelectric Phase Transitions in Inorganic Materials, talk presented in Session B (June 26), Annual Meeting American Crystallographic Association, Philadelphia, June 26–July 1, 1988. [This talk was based in part on Structurally-Based Prediction of Ferroelectricity in Inorganic Materials with Point Group *6mm*, Acta Crystallographica, **B44** (December 1988)].

[12] Siegrist, T., Applications of the Crystallographic Search and Analysis System CRYSTDAT in Materials Science, J. Res. Natl. Inst. Stand. Tech. 94, 1 (1989).

[13] Allen, F. H., Bergerhoff, G., and Sievers, R., (Editors), *Crystallographic Databases—Information Content, Software Systems, Scientific Applications*, Data Commission of the International Union of Crystallography, Bonn, W. Germany (1987).

Volume 94, Number 1, January-February 1989
Journal of Research of the National Institute of Standards and Technology

NIST/Sandia/ICDD Electron Diffraction Database: A Database for Phase Identification by Electron Diffraction

Volume 94 Number 1 January-February 1989

M. J. Carr and W. F. Chambers

Sandia National Laboratories
Albuquerque, NM 87185

D. Melgaard

J&M Systems
Albuquerque, NM 87123

V. L. Himes, J. K. Stalick, and A. D. Mighell

National Institute of Standards and Technology
Gaithersburg, MD 20899

A new database containing crystallographic and chemical information designed especially for application to electron diffraction search/match and related problems has been developed. The new database was derived from two well-established x-ray diffraction databases, the JCPDS Powder Diffraction File and NBS CRYSTAL DATA, and incorporates 2 years of experience with an earlier version. It contains 71,142 entries, with space group and unit cell data for 59,612 of those. Unit cell and space group information were used, where available, to calculate patterns consisting of all allowed reflections with d-spacings greater than 0.8 Å for ~59,000 of the entries. Calculated patterns are used in the database in preference to experimental x-ray data when both are available, since experimental x-ray data sometimes omits high d-spacing data which falls at low diffraction angles. Intensity data are not given when calculated spacings are used. A search scheme using chemistry and r-spacing (reciprocal d-spacing) has been developed. Other potentially searchable data in this new database include space group, Pearson symbol, unit cell edge lengths, reduced cell edge length, and reduced cell volume. Compound and/or mineral names, formulas, and journal references are included in the output, as well as pointers to corresponding entries in NBS CRYSTAL DATA and the Powder Diffraction File where more complete information may be obtained. Atom positions are not given. Rudimentary search software has been written to implement a chemistry and r-spacing bit map search. With typical data, a full search through ~71,000 compounds takes 10~20 seconds on a PDP 11/23-RL02 system.

Key words: electron diffraction; identification; numeric database; phase characterization.

Introduction

The identification of crystalline objects in the size range from 10 μm to 10 Å is readily accomplished in the analytical electron microscope (AEM) if the analyst has access to appropriate information. Most often the needed information exsists, but either it is not readily accessible in the laboratory or it is not in the most useful form. Acquiring and reprocessing reference data is often the time-limiting step in the identification process. Information scattered through the open literature has been collected into compilations which recently have become available in computer-readable form [1,2]. Even so, the format of the data is not ideally suited for electron diffraction work [3].

We perceived a need for a specialized database to support efficient phase identification by combined electron diffraction and energy dispersive x-ray spectroscopy (EDS) in a modern analytical electron microscope. Considering the quality of the experimental data obtainable from the AEM, the quantity of reference data, and available computing machinery, we set out to create a database to sup-

port search/match procedures [4] and crystallographic calculations [5] performed routinely in our laboratories.

Description of the Database

This database was derived from two copyrighted databases, NBS CRYSTAL DATA and the PDF-2. The preparation of the derivative database was facilitated by the fact that the original databases are in the same format since both were built with a program called NBS*AIDS83 [6]. The new derivative database contains a subset of information from the full databases, selected on the basis of pertinence to electron diffraction analysis. Only inorganic compounds were used [7]. The data is accurate and as complete as possible, but has been reduced in precision to a level appropriate for electron diffraction work ($\pm\sim 1\%$@ 1.5 Å). It has been packed in a manner which allows it to be used on a small computer equipped with a 10 Mb hard disk. The database is complete so that it is useful without reference to other sources such as cards [1] or books [1,2], but it contains pointers so that if a card file [1], CDROM [8], or other full listing [1,2] is available, one can quickly get to that information as well.

The data were selected from the two sources as follows:

1. All inorganic compounds from NBS CRYSTAL DATA were used. The unit cell and space group information from each compound were used to compute up to 60 non-redundant allowed reflections with d-spacings greater than 0.8 Å. Intensities were not computed. There are 59,612 entries of this type.
2. Inorganic compounds from PDF-2 sets 1-33 whose entries do not give unit cell data, and all entries from sets 34-36 were used. These are only a subset of the full PDF-2 database. It was assumed that entries having unit cell information in sets 1-33 are adequately represented by similar entries in NBS CRYSTAL DATA and would only duplicate information. d-spacings and intensities (obtained from x-ray methods) were used. All inorganic compounds from PDF-2 sets 34-36 were used whether or not they contained unit cell information, since it could not be assessed whether such compounds had been included in NBS CRYSTAL DATA yet. (A little duplication is better than missing a compound altogether.) This group contains 11,530 entries.

Despite their different origins, the two types of source data are functionally equivalent and are treated equally in the new database. They are mingled in the ordered and indexed Search file. The computed data (1.) represents the best target group for matching on the basis of observed d-spacings from single diffraction patterns. The data in group (2.) is similar to the data obtainable from the PDF Level I database, an earlier version of the PDF-2 used in this work. We have searched against type (2.) data for over two years with fair success [3]. When searching failed, it was often because the experimental x-ray observations in the PDF Level I database did not include high d-spacing reflections observable by electron diffraction. The computed data in (1.) is an attempt to correct this weakness, but computation is not possible for compounds in (2.) because unit cell and space group information is absent. The data in (2.) is valuable nonetheless, because even if you cannot completely characterize such a material, at least you can determine that "you found it again." The literature reference may be of some use in such cases.

As in the earlier version of this work, data are stored in two types of files: Reference files and a Search file. We have kept sufficient information in each entry to be of use for electron diffraction analysis, but have put only certain critical information in the Search file, for the sake of speed. The data for each compound, therefore, is divided between the Search file and a Reference file. There may be more than one entry for a given compound. Multiple entries for the same compound are present mainly when derived from different literature citations.

The contents of a Reference file entry for a given compound are:

1. Name length (1 byte)	Number of bytes (x) to store the compound name.
2. Formula length (1 byte)	Number of bytes (y) to store the compound formula.
3. # of intensities (1 byte)	Number of reflections (z) having intensities (if computed, then 0).

4. Unit cell angles (4 bytes)	Number and kind of angles given for the conventional unit cell.
5. Reduced cell angles (4 bytes)	Number and kind of angles given for the reduced cell [4].
6. Pearson symbol (4 bytes)	xXnnnn, indicates crystal class, symmetry, and number of atoms in the conventional unit cell.
7. Journal reference (17 bytes)	CODEN, volume, page, and first 9 characters of the author name field (Radix-50), and year (-1800).
8. Source ID (3 bytes)	PDF number or CRYSTAL DATA ID number.
9. Unit cell angles (0–12 bytes)	Degrees*100, only the necessary ones.
10. Reduced cell angles (0–12 bytes)	Degrees*100, only the necessary ones.
11. Compound name (x bytes)	Including mineral name, if applicable (Radix 50).
12. Compound formula (y bytes)	Functional formula (ASCII).
13. Intensitites ($\geqslant z$/2bytes)	In nibbles, if present (a nibble=4 bits=1/2 a byte) always ending on a word boundary

The first eight items are fixed length fields; the last five vary in length and may be absent. The entries in the Reference files therefore vary in length. Angles are multiplied by 100 and rounded to convert them to integers, which take less storage than floating point numbers while preserving two decimal place precision. Only angles not equal to 90 degrees are stored, with a code indicating whether they represent α, β, or γ. Missing angles are always 90°. The compound names are converted to Radix-50 notation which encodes 3 characters per 2 bytes (50% denser than packed character strings). Reference entries are grouped together in 16 Reference files, each of which contain a large number (2000-5000) entries. Twelve reference files contain data from NBS CRYSTAL DATA entries, and four files contain data from PDF-2 compounds. There is a pointer to the corresponding Reference file entry stored in each Search file entry. The Reference files are not meant to be searched, but rather to be directly accessed one time, after a search has been completed. In total, these files require ~4.5 Mb of disk.

The contents of a Search file entry for a given compound are:

1. Chemistry bit map (12 bytes):	Elements 1-96 in six 16-bit words.
2. r-spacing bit map (22 bytes):	Eleven 16-bit words, representing 176 cells, each 0.018 Å wide, of r-spacing ($r=\lambda L/d$, where $\lambda L=2.5$ Å-cm). At λL=2.5 Å-cm, r-spacings range from 0.0 to 3.2 cm, representing d-spacings from ∞ to 0.8 Å).
3. Space group (2 bytes):	Encoded in two bytes, allows for *nnnX *, if present, signifies that the space group is not completely determined, so an aspect is given [6]. nnn is the space group number (1-230). X, if present, gives the setting, e.g., Pbca or Pcab, etc.
4. Unit cell edges (6 bytes):	Å*100, three 16-bit integers, the dimensions of the unit cell given in NBS CRYSTAL DATA, which may be different from the unit cell assigned by the original author.

5. Reduced cell edges (6 bytes): Å*100, three 16-bit integers, the dimensions of a mathematically unique primitive unit cell equivalent to the conventional unit cell.
6. Reduced cell volume (2 bytes): Used by the NBS Lattice search program.
7. Flags (8 bits): Organic/inorganic, mineral, metals & alloys, hydrate, deleted, NH_x-containing, unit cell differs from original author's cell, and a spare.
8. Pointer (3 bytes): To the corresponding Reference file entry.
9. Spare word (2 bytes): In case something simple needs to be added in the future.

This is a single large file (~4 Mb). The entries are ordered on the basis of composition, beginning with atomic number 11 (sodium). We have assumed an EDS detector with a beryllium window is being used. This is the most common type of detector in the field today. It is capable of detecting only elements whose characteristic x rays are hard enough to penetrate the Be window (namely $Z \geqslant 11$). This orders the file on the basis of EDS-detectable qualitative chemistry, scattering oxides, carbides, etc., through the file associated with their EDS-observable elements. This ordering is advantageous even when using an EDS detector that can detect lighter elements, because the light elements are so common in compounds in the file as to be a disadvantage when searching. For example, oxygen is present in more than half of all the compounds in the file, so it is much more efficient to go looking for iron-bearing compounds (5909) that contain oxygen (3837), than oxygen-bearing compounds (40084) that contain iron (3837). The ordering scheme also places compounds containing only undetectable light elements (e.g., ice, graphite, boron nitride) at the end of the file, where they may be skipped as a group if so desired. Each entry in this file is a fixed length (56 bytes). Entries are grouped into records. There are 18 entries in each record, followed by 16 empty bytes to pad the record length to 1024 bytes (two blocks). This facilitates a speedy search by creating a constant offset or spacing between fields of the same type within a record, and allows for easy disk access with a two-block buffer.

The first part of the Search file contains an index to the records in the remainder of the Search file. The indexing scheme was described in detail previously [3].

There is one index entry for each record in the Search file. The Index file is 60 Kb in size. There are 18 compounds per 1024-byte record in the Search file, so each entry in this index file refers to 18 compounds. Because the Search file is ordered by chemisty, the Index file makes it possible to perform a coarse screening (in groups of 18) of the Search file to find the records which may contain compounds with the proper chemistry. More directly, the index allows the search software to apply a quick test and then most often skip over a group of compounds which certainly contain no possible matches based on the chemistry requirements. This greatly reduces the number of Search file entries which must be processed in detail and can increase the overall speed of the search by as much as an order of magnitude.

The structure of the file is based on our search/match experience with an earlier version of this database. It is designed to be searched first on the basis of chemistry, which has been shown to be the primary characteristic in electron diffraction phase identification work [9]. The index file allows one to skip over large sections of the file where no chemistry matches are possible, greatly reducing search time. After considering chemistry, we can perform a secondary match on the basis of observed r-spacings, or on the basis of flags indicating membership in one or more subsets of the data. It is also possible to make no requirements on chemistry, in which case all entries in the file will pass the chemistry test. Then, a search takes the maximum amount of time since every entry will be tested for secondary match requirements. It is possible to search on the basis of other parameters, such as space group, Pearson Symbol, reduced cell parameters [4], or unit cell parameters, although we have not devel-

The contents of an Index entry for a given compound are:

1. Bnum: Block number in the Search file (2 bytes).
2. ORmap: Six 16-bit words containing the result of performing the boolean OR function on the chemistry bit maps for all the compounds in one record.

oped software to do so. Since the unit cell parameters are stored for most of the compounds, it is possible to write additional software to quickly calculate precise d-spacings and Miller indices of allowed reflections for a particular compound if the need arises.

Search/Match Software

Source code for basic functional search/match software is distributed with the database. Two versions exist. An assembly language search algorithm was written and described for the first generation of this file [3]. The general nature of the algorithm remains the same for this file, with minor changes to accommodate the format of the new database. Experience with typical data (2 or 3 observed elements, all unobserved heavy elements and some light elements excluded, 6-8 diffraction spots) has shown that most searches require 10-20 seconds to search the full file on a PDP 11/23 equipped with an RL02 10 Mbyte hard disk; I/O takes several times longer than that. It is also possible to write search programs for this file in high level languages. FORTRAN versions have implemented the same search on VAX, PDP, and SUN computers. On the PDP, the FORTRAN version gives the same results but runs five times slower than the assembly language version. Similar programs could be written in other languages that support bit manipulation. A version of this software has been written in Flextran to be integrated into the RAD group of programs [5] which run on computerized EDS analysis equipment attached to an electron microscope. Users are encouraged to modify or add to the programs. Additional software for searching on the basis of reduced cell [4] or space group may be added at a later date.

Conclusion

The database described in this report contains what we believe to be the only complete collection of inorganic compound data structured for phase identification by electron diffraction available. Nevertheless, the database is small enough to reside on a personal computer or laboratory microcomputer dedicated to EDS analysis in an electron microscope laboratory.

Since the database was designed especially for electron diffraction analysis, it is not expected to work well for traditional x-ray diffraction analysis where high precision data for both peak position and intensity are obtained and used.

It is anticipated that many different search/match schemes will be able to use this database, although we have initially implemented only one. Searching first on the basis of qualitative EDS chemistry is a natural consequence of the type of information obtained with the AEM and greatly increases searching speed in a small computer. The computed data, incorporating high d-spacing reflections, are very diagnostic for electron diffraction search/match identification. Beyond its usefulness as a search/match tool, the database also provides a convenient resource for crystallographic data for pattern simulation. The full integration of this database into our existing analytical software is planned, and we expect that it will be useful to other laboratories as well.

The development of this database has been a joint project at Sandia National Laboratories and the National Institute of Standards and Technology, with the encouragement of the JCPDS/ICDD. Further evolution of this database and any related items will be guided by the members of the Phase Identification by Electron Diffraction subcommittee of the JCPDS Technical Committee and the NIST Crystal Data Center. The details of the database format, software to generate and update the database from original source tapes, and the search/match software are available upon request. The database itself is copyrighted by the National Institute of Standards and Technology and is being distributed by license through the JCPDS/ICDD. For information on obtaining the database contact JCPDS/ICDD headquarters [1] or the NIST Crystal Data Center [2].

References

[1] PDF-2. The master database of chemical, crystallographic, and x-ray powder diffraction data compiled and evaluated by the Joint Committee on Powder Diffraction Standards, International Centre for Diffraction Data, 1601 Park Lane, Swarthmore, PA, 19081. The PDF-2 database contains all the information on the familiar PDF cards and is available on cards and magnetic tape by license only.

[2] NBS CRYSTAL DATA (1987). The master database of chemical and crystallographic data compiled and evaluated by the NIST Crystal Data Center, National Institute of Standards and Technology, Gaithersburg, MD 20899. The full database is available on magnetic tape by license only. Portions of the data are available in book form as Crystal Data Determinative Tables in several volumes.

[3] Carr, M. J., Chambers, W. F., and Melgaard, D., A Search/Match Procedure for Electron Diffraction Data Based on Pattern Matching in Binary Bit Maps, Powder Diffraction **1**, 226 (1986).

[4] Himes, V. L., and Mighell, A. D., NBS*LATTICE, A Program to Analyze Lattice Relationships, NBS Technical Note 1214, 1985, NIST Crystal Data Center, National Institute of Standards and Technology, Reactor Radiation Division, Gaithersburg, MD 20899.

[5] Carr, M. J., and Chambers, W. F., A Review of Crystallographic Calculational Methods used in the RAD Group of Computer Programs for Analytical Electron Microscopy, J. Microsc. **134**, pt. 1, 55 (1984).

[6] The NBS*AIDS83 data evaluation and database-building computer program was developed at the National Institute of Standards and Technology for use by both the JCPDS/ICDD and the NIST Crystal Data Center. Two distribution databases were created and are maintained with this program—NBS CRYSTAL DATA and JCPDS/ICDD PDF-2. The research uses of this program have been described earlier (Mighell, A. D., Hubbard, C. R., and Stalick, J. K., NBS*AIDS80: A FORTRAN Program for Crystallographic Data Evaluation, NBS Technical Note 1141, 1981). The user is referred to a description of NBS CRYSTAL DATA for detailed information on many items in common to both databases (Stalick, J. K., and Mighell, A. D., CRYSTAL DATA. Version 1.0 Database Specifications, NBS Technical Note 1229, 1986). The program as implemented by the JCPDS/ICDD contains many codes and conventions specific to the PDF-2 database and descriptions of these items are included in a program user's manual which is available from the JCPDS/ICDD or the NIST Crystal Data Center.

[7] Organic compounds, representing more than 80,000 additional entries, were not included because, in general, they are sensitive to degradation in the electron-beam/high-vacuum environment and are rarely successfully analyzed by AEM. On a day-to-day basis in most materials science laboratories where this database is likely to be used, these compounds would only occupy disk space and increase search times. If a need arises in the future, these compounds could be added to this database with no modification of its basic structure. Some compounds with organic components, flagged as inorganic in the original databases, are included in the new database.

[8] The PDF-2 and NBS CRYSTAL DATA, which in full AIDS format occupy ~250 Mbyte of storage, have recently become available on an optical, read-only mass storage device called a CDROM, which is similar to CDs available for home audio systems. A drive which reads the CDROM is available for IBM and compatible PCs. Such devices are relatively inexpensive, and are expected to become commonplace in laboratories using this type of data.

[9] Anderson, R., and Johnson, G. G., The Max-D Alphabetical Index to the JCPDS Database: A New Tool for Electron Diffraction Analysis, ed. Bailey, G. W., *37th Annual Proceedings of the Electron Microscopy Society of America*, San Antonio, Texas, (1979) p. 444.

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

Numeric Databases in Chemical Thermodynamics at the National Institute of Standards and Technology

Volume 94 Number 1 January–February 1989

Malcolm W. Chase

Chemical Thermodynamics Division
National Institute of Standards and Technology
Gaithersburg, MD 20899

During the past year the activities of the Chemical Thermodynamics Data Center and the JANAF Thermochemical Tables project have been combined to obtain an extensive collection of thermodynamic information for many chemical species, including the elements. Currently available are extensive bibliographic collections and data files of heat capacity, enthalpy, vapor pressure, phase transitions, etc. Future plans related to materials science are to improve the metallic oxide temperature dependent tabulations, upgrade the recommended values periodically, and maintain the bibliographic citations and the thermochemical data current. The recommended thermochemical information is maintained on-line, and tied to the calculational routines within the data center. Recent thermodynamic evaluations on the elements and oxides will be discussed, as well as studies in related activities at NIST.

Key words: chemical thermodynamics; chemical thermodynamics data; data files; enthalpy; evaluated data; heat capacity; JANAF Thermochemical Tables; numeric databases; phase transitions; vapor pressure.

Accepted: December 1, 1988

Introduction

The Chemical Thermodynamics Data Center provides the chemical process industry with critically evaluated thermodynamically consistent data which can be used to establish the equilibrium constants and enthalpies of reaction for important chemical reactions. These critically evaluated data are used in the design and interpretation of research in physics, chemistry, biochemistry, geochemistry, environmental sciences, metallurgy, and other fields where chemical interpretations are important. The center provides this data describing the change in the chemical properties of substances as well as bibliographic reference services in thermochemistry.

In characterizing the thermodynamic properties of chemical species, the primary effort involves the study of inorganic species and their aqueous solutions. Additional efforts are also directed at the examination of organic species and biological systems. In the area of inorganic chemistry, this data center has the responsibility for two major publications: The NBS Tables of Chemical Thermodynamic Properties [1] and the JANAF Thermochemical Tables [2]. The former publication deals with the property values at 298.15 K for 92 elements and their compounds while the latter presents temperature-dependent values for 49 elements and some of their compounds.

Currently, the data center staff is involved in extending and upgrading the information contained in these two publications. Although the projects which led to these publications as a product are continuing, they are following a different approach from that used in the past. Part of this change re-

sults from the merger of the two above-mentioned projects. In addition there are many cooperative ventures in which the data center is involved. These also require a slightly different approach to the evaluation process. The following discussion will mention the accessibility of the NIST recommended data and the procedures being followed for the study of the elements and the oxides.

Accessibility of Thermodynamic Data

The information contained in the previously mentioned two publications is easily available to users in hard copy. Both publications appeared as supplements to the Journal of Physical and Chemical Reference Data and are readily available from the American Chemical Society. Additionally, magnetic tape versions of both publications are available from the Office of Standard Reference Data at the National Institute of Standards and Technology. Both data sources are also available through an on-line commercial vendor, CAS-STN. The databases are called NBSTHERMO and JANAF, respectively. Thus, in principle, this thermodynamic information is quickly and easily obtainable for any users.

Within the NIST Chemical Thermodynamics Data Center, these resources are accessible via an on-line system. This has been developed to tie together the information from these databases with many of the programs used in the critical evaluation of data. Current data evaluation efforts are in progress for the JANAF Thermochemical Tables, the NBS Tables of Chemical Thermodynamic Properties, the CODATA Task Group on Chemical Thermodynamic Tables, as well as other cooperative efforts. It is necessary that each member of our staff have access to the centralized collection of this recommended information.

The NBS Tables are stored and retrieved on our HP-1000 system[1] using commercially available database management software, IMAGE/1000 and ASK/1000. The database is searchable via the chemical formula, as it appears in the published hard copy [1]. The retrieval process provides up to six thermodynamic property values, namely, enthalpy of formation at $T/K=0$ and 298.15, and heat capacity, entropy, enthalpy, and Gibbs energy of formation, all at $T/K=298.15$.

The JANAF Thermochemical Tables are stored and retrieved using a locally developed search and retrieval system called SETKY-GETKY [3]. Searching the database is done via chemical formula (in the normal order) or chemical formula (in the Chemical Abstracts order) coupled with the physical state. The search returns the temperature-dependent thermal functions, as published in the hard copy [2]. Additional information includes the enthalpy of formation at 298.15 or 0 K, the value of the gas constant used, reference temperatures and pressures, transition temperatures, and origin and date of the table.

The main points to remember are that (1) the thermodynamic information for each species is obtainable during an interactive session at the computer terminal, (2) the thermodynamic information is centrally located on our minicomputer, and (3) the data evaluation programs used in our data center are tied into these databases programmatically.

Element Study

Our efforts in the characterization of the thermodynamic properties of the elements illustrate our current evaluation approach. For our purposes the study of the elements involves the condensed phases (including all crystal, liquid, and amorphous states) and the gaseous phases (including the monatomic gas), any additional gas phase n-mers, and any positive and negative ions that exist.

The characterization of the gas phase molecules is very important in that they are necessary for the description of the vapor phase above any condensed phase elemental system. As an example, near the melting point of boron, the ratio of the vapor pressure of $B_2(g)$ to $B(g)$ is approximately 0.000009. Thus, at this temperature or lower, the contribution of the dimer to the total vapor pressure is unimportant. However, in the case of sulfur at the normal boiling point, approximately 717 K, at least eight species exist in the vapor, with $S_6(g)$, $S_7(g)$, and $S_8(g)$ being the dominant species, contributing 95% of the vapor pressure at this temperature.

[1] Certain commercial equipment including computer software is identified in this paper in order to adequately specify procedures, experiments, and techniques used. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials, equipment, or software identified are necessarily the best available for the purpose.

Oxide Study

The same approach is used in the study of the oxides. We are concerned with the characterization of the condensed and gaseous phases, including the examination of the possibility of the existence of *n*-mers in the gas phase. Currently under examination are the alkaline earch metal oxides and the transition metal oxides (specifically, iron, nickel, and cobalt).

Process for Evaluating Data

In general, the goal of the study of the thermodynamic properties of the elements is to generate temperature-dependent values for the heat capacity, enthalpy, Gibbs energy function, entropy, enthalpy of formation, Gibbs energy of formation, and the log of the equilibrium constant of formation. These tables also include various first-order transtition processes, such as solid-solid, fusion, sublimation, vaporization, etc. Second-order transitions must also be included, such as Curie points and superconducting temperatures.

For both the elements and the oxides, the evaluations are carried out using a multistep approach, with each step yielding a documented publication. Our users should be able to follow our work easily and see it on a shorter timetable.

The first step in this approach is to generate an annotated bibliography. The literature is constantly being scanned for data pertaining to the thermodynamic properties of the elements and the oxides. Relevant information is entered into bibliographic files, which are revised whenever new references are encountered. Each literature citation contains the following information:

1. author names
2. article title
3. journal citation
4. abstracting service reference
5. annotation indicating property studied

When a literature citation is found, the main emphasis is to enter the author names and journal citation. The remaining information will be added at a later time, if it is not readily available. The prime intent is to produce a complete listing of literature citations which is available to any user immediately. The remaining information, although useful and necessary for the final evaluation process, can be added later. Since these files are actively maintained on our computer, current versions may be printed immediately upon request.

In the second step, the relevant data are collected into tables and graphs, as references become available, and are stored in disk files actively maintained on our computer system. A combination of the bibliographic and numerical information is gathered together into a series of data summaries for easy appraisal of the quantity and quality of data available. At this point, the data (as stored in the disk files) can be directed through a series of programs to produce appropriate graphs which indicate visually the qualitative agreement (or the lack thereof) among the various temperature-dependent data sets. In addition, tabular summaries of the available studies are also maintained. At this point, no statement is made as to the reliability or goodness of any study. The intent is to know the studies which exist and what data are derived from them. The data summaries can also be distributed to interested users.

These "collection" phases are followed by an "evaluation" phase in which the available data are intercompared and, where possible, assessed against theory. In this phase, the recommendations are prepared. The process is described below for condensed phases and gaseous species.

Condensed Phases

The generation of the thermal functions for the condensed phases of the elements and the oxides (that is, the heat capacity, enthalpy, entropy, and the Gibbs energy function) is based on a numerical integration of adopted or recommended heat capacity values. In general, these recommended heat capacity values are derived from a detailed analysis of many sources.

In order to accomplish this analysis, the literature search is for these prime sources of data: all heat capacity and enthalpy studies as a function of temperature—effects which cause changes in the heat capacity and discontinuities in the enthalpy must be included in the study. These include effects such as structural transitions, Curie temperatures, superconductive temperatures, and, in general, any first- or second-order transition.

After collecting the available literature, the data are extracted and summarized. The heat capacity and enthalpy are amenable to graphical comparisons. Typically there are four variations of graphs which are useful to consider:

1. C_p vs T (all T)
2. C_p/T vs T^2 (for $T<30$ K), mainly for elements
3. $(C_p-\alpha T)/T^3$ vs T (for $T<30$ K), mainly for elements
4. $H(T)-H$ (298.15 K)$/(T-298.15)$ (for $T>298.15$ K)

For the other data, such as temperatures and enthalpies of fusion, tables of the various studies are maintained for comparison purposes.

Currently, the emphasis is directed towards the literature surveys, the data collection process, and the summary and display of the temperature-dependent information. Of course, data summaries are maintained for other information, such as the transition temperatures and enthalpies.

Gas Phase Species

For the monatomic gases spectroscopic information is necessary for the evaluation and generation of thermochemical tables. Most important is the knowledge of the atomic energy levels, especially the low lying levels, the ionization potential, and the electron affinity. The atomic energy level information is obtained from the NIST Data Center on Atomic Energy Levels and is stored on-line so that calculations can be done at any time using a variety of calculational procedures.

For the n-mers, the information which is normally retrieved is the spectroscopic information dealing with the electronic energy levels and vibrational-rotational energy level information. Again, this information is maintained in data summary tables for easy comparison. As implied in the calculation of the thermodynamic properties for monatomic gases, the extent and type of data available for the n-mers will really determine the calculational pathway used. Significant difference can occur.

Also important is the vapor pressure information relating to sublimation, vaporization, decomposition, and reaction processes. Here, a plot of log p vs $1/T$ would be made. In this case, plots are useful not only to show the vapor pressure results but also to confirm the specification of the units (unfortunately the units are not always given in the literature). As with other data studies, a tabular listing of the available studies is also maintained.

Previous Critical Evaluations

Also important in the collection of data is the listing of the various critical evaluations that are already in existence. For example, if your interest is in aluminum, there are five to be considered! In most cases, the most recent evaluation would be preferred, since more data would be available. But there are cases in which evaluations differ. It is not always clear as to the cause of these differences, but they normally imply problems in the interpretation of the existing data.

Summary

The prime recommended thermodynamic information generated within the Chemical Thermodynamic Data Center at NIST is available in hard copy from the American Chemical Society, in magnetic copy from the Office of Standard Reference Data at NIST, and interactively via our data center computer and CAS-STN.

As for the characterization of the thermodynamic properties of the elements and their oxides, the information is being made available in a series of sequential publications which will present annotated bibliographies, detailed data summaries, and evaluated thermochemical tables. During this sequential process, the same information is available through a series of active computer files. For those who need to be current as to new developments in the elemental and oxide thermodynamic information, we should have available annotated bibliographies, data summaries in terms of graphs and tables, and listings of available critical evaluations. More importantly, we can provide you with a detailed thermochemical table for the chemical species of interest. Currently, we have the ability to provide this information for roughly 50 elements and their oxides.

References

[1] Wagman, D. D., Evans, W. H., Parker, V. B., Schumm, R. H., Halow, I., Bailey, S. M., Churney, K. L., and Nuttall, R. L., The NBS tables of chemical thermodynamic properties. Selected values for inorganic and C1 and C2 organic substances in SI units, J. Phys. Chem. Ref. Data **11**, Suppl. 2, 1982.

[2] Chase, M. W., Davies, C. A., Downey, J. R., Frurip, D. J., McDonald, R. A., and Syverud, A. N., JANAF Thermochemical Tables. Third Edition, J. Phys. Chem. Ref. Data **14**, Suppl. 1, 1985.

[3] Bickham, D., and Neumann, D., User's Guide for SETKY-GETKY. A Keyed Access System for the HP1000, NBSIR 86-3417, October 1986.

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

Numeric Databases for Chemical Analysis

Volume 94 Number 1 January–February 1989

Sharon G. Lias

National Institute of Standards and Technology
Gaithersburg, MD 20899

Databases for use with analytical chemistry instrumental techniques are surveyed, with attention to existing databases and collection efforts now underway, as well as needs for new databases. Collections of spectra for use in NMR, infrared spectroscopy, and mass spectroscopy are described. Using mass spectral databases as an example, a critique is presented of automated quality control procedures used to evaluate individual spectra in large collections; the kinds of problems which have been encountered in using these procedures are discussed. Finally, a brief critical review is presented covering the application of computers to the identification of unknown compounds using spectral databases; again, algorithms used with mass spectrometry are taken as the example. Ongoing work at NIST with the NIST/EPA/MSDC Mass Spectral Database is concerned with many of these problems; recent developments are described.

Key words: analytical chemistry; computer; database; evaluation; infrared spectrum; mass spectrum; nuclear magnetic resonance.

1. Introduction

In principle, the measurement technique in which spectroscopy is used as an analytical tool involves obtaining a spectrum of the sample of interest (the "unknown") and identifying the unknown compound by the similarity of its spectrum to that of a particular ("known") chemical compound. Here we use the word "spectroscopy" in the broadest possible sense; spectroscopy is taken to be any experimental technique which provides a reproducible "spectrum" characteristic of particular chemical species. This includes, for example, all optical spectroscopy, nuclear magnetic resonance, electron spin resonance, mass spectrometry, and so on.

Of course, from the beginning of the use of spectral techniques in the analytical laboratory, it was recognized that the comparison spectra need not be obtained at the same time, or even on the same instrument, as the analysis itself. Because one could collect standard spectra and use them over and over again, it is not unexpected to find that there is a long history of data collection efforts aimed at analytical applications [1,2]. With the beginning of the computer age, it was of course a natural extension of these activities to store spectral databases on computers, and to conduct automated searches of those databases in order to "match" the spectrum of the unknown compound with that of a standard reference compound. The use of automated instruments equipped with reference libraries has become a well-established measurement technique for analytical chemistry. At the present time, computerized algorithms are also used to evaluate the large numbers of spectra which comprise these collections.

In spite of the long history of data collection efforts involving analytical spectra, there is some dissatisfaction with the size and quality of available collections. For example, in 1986 Thomas L. Isenhour wrote an editorial [3] describing the consensus of experts concerning computerized databases for use in analytical chemistry measurement techniques: "... the current state of spectroscopic databases is such that it inhibits good applications of known search and interpretive procedures as well as further research on these methods. ... we do not in general have high quality spectroscopic databases available... Perhaps 10 million chemical compounds are now known. Some measurements have been made on all of them. Very few, if any, structure identifications have been made in recent times without resorting to some form of spectroscopy. Why then are the largest available spectral data files in computer format limited to a few tens of thousands of compounds?"

This paper presents a brief survey of the use of automated databases as an integral part of spectroscopic measurement techniques for analytical chemistry, with emphasis on mass spectrometric databases. Because of the rather dim view by the experts of the analytical databases in common use, the survey includes a list of the most popular automated analytical databases with attention to the numbers of spectra available in each of them. A discussion of the current state of automated evaluation algorithms being used with mass spectral databases is included.

Ongoing work aimed at updating and improving the quality of the mass spectrometric database distributed by the National Institute of Standards and Technology Office of Standard Reference Data is described.

2. Brief Survey of Automated Analytical Databases

2.1 Nuclear Magnetic Resonance Spectroscopy (NMR)

The databases listed below are all provided with software which enables the user to look up particular spectra or to match the characteristics of a particular spectrum of an unknown compound. Most NMR databases also include software for spectrum estimation and interpretation.

2.1.1 C-13 NMR Database on the Chemical Information System The Chemical Information System [4] collection currently consists of a total of 11,700 ^{13}C NMR spectra. The database was last updated in November 1985, when many incorrect assignments in older spectra were corrected, and over 4,000 new spectra were added. The database was originally put together by the Royal Dutch Chemical Society (also called Netherlands Information Combine).

2.1.2 C-13 NMR Online Service of the Fachinformationszentrum (FIZ), Karlsruhe, W. Germany (accessed in the U.S. through STN International) This widely-used NMR database was added to the STN system [5] in December 1987, having been marketed previously in the U.S. by Scientific Information Service (SIS). The collection contains 67,500 ^{13}C chemical shifts, coupling constants, and relaxation times.

2.1.3 Bruker Spectroscopic Database This database is available to Bruker customers for on-site use. It requires a Bruker Aspect 2000 or 3000 computer together with a Bruker software package (BASIS—Bruker Automatic Spectroscopy Interpretation System). The database contains various modules, including ^{13}C NMR (19,000 spectra), ^{1}H NMR (900 spectra), as well as a combined ^{13}NMR-MS database.

2.1.4 Sadtler Laboratories This database consists of 24,000 sets of ^{13}C NMR chemical shifts with compound names, and also 10,000 ^{13}C NMR spectra in full digital format that can be used to view expanded displays of the spectra [1c]. The database is designed for use with Sadtler's own ^{13}C search software package, which operates on IBM-compatible personal computers.

2.1.5 Collection of National Chemical Laboratory for Industry, Japan The integrated online "Spectrum Database System" [6], which includes collections of NMR, ESR, IR, Raman, and mass spectra has both ^{1}H NMR spectra (6,000 compounds) and ^{13}C NMR spectra (5,700 compounds) along with search software enabling a user to look up a particular spectrum (and conditions under which it was run) or to match an unknown spectrum. All spectra were determined at the NCLI under carefully controlled conditions.

2.1.6 Other Collections of NMR Spectra The list given above is not exhaustive. For example, Varian also markets an NMR database, and Tsukuba University (Japan) produces a CD-ROM collection of ^{13}C NMR spectra of polymers. The data came from existing handbooks. The system also contains programs to synthesize the NMR spectra from structural information.

2.2 Infrared Spectra (IR)

In the field of infrared spectroscopy, many large collections of spectra were built up [1,7] at a time when the spectrometers in use were prism and grating instruments. Within the past decade, the instrumentation in general use in analytical laboratories has changed to Fourier transform infrared spectrometers (FT-IR), which generate digitized spectra. Although the older analogue spectra can be digitized to be made compatible with the data systems of the newer instruments, questions have been raised about the desirability of doing this. In the opinion of some experts [8], many of the older collections of spectra are no longer adequate to serve as reference spectra for comparison with results taken on the newer instruments. For this reason, effort has been given recently to building completely new collections of IR spectra which were generated in digital format in FT-IR instruments. In the discussion which follows, attempts will be made to distinguish between the newer digitized collections, and databases of spectra from prism and grating spectrometers.

2.2.1 Aldrich-Nicolet Digital FT-IR Database and the Sigma-Nicolet Biochemical Library Nicolet, in collaboration with Aldrich and with Sigma, is producing high quality databases of FT-IR spectra of the compounds in the catalogues of these two companies. The Aldrich-Nicolet collection contained 10,600 compounds and the Sigma-Nicolet collection, 10,400 compounds in 1987. These databases are being updated in 1988 with the addition of several thousand new spectra. The databases are designed for use on several popular personal computers, and are distributed with software which is geared to locating spectra which match the peak intensities and locations from an IR spectrum of an unknown substance.

2.2.2 Sadtler Research Laboratories Spectra The largest commercially available collection of infrared spectra [1c], with >60,000 spectra largely from prism and grating spectrometers. The current collection also includes some FT-IR spectra.

2.2.3 Coblentz Society Spectra Beginning in the mid-1960s, the Coblentz Society, in collaboration with the Joint Committee on Atomic and Molecular Physical Data (JCAMP), put together a collection of 10,500 donated infrared spectra taken on prism and grating spectrometers. The effort included developing evaluation procedures for IR spectra, and evaluating the entire collection of spectra. The collection was originally distributed in 10 volumes in a looseleaf notebook format [7]. Recently, 4,400 of these spectra have been digitized, and will be made available through the Coblentz Society, which is also digitizing the remaining spectra. Dr. Clara Craver, of the Chemir Labs, who played a key role in putting together the original Coblentz Society collection, is actively soliciting donations of new spectra to increase the size of the database, which will be available in a format for use with personal computers.

2.2.4 EPA Vapor Phase Spectra This collection of 3,300 spectra originated in laboratories of the EPA, and is in the public domain. Although not commercially available as a collection, the spectra are available through the instrument companies manufacturing IR spectrometers.

2.2.5 Collection of the Univeristy of California-Riverside "Clearinghouse for Digital Infrared Spectra" A new project was initiated in October 1986 for the collection of a database of digitized FT-IR spectra under the leadership of Drs. Peter Griffiths and Charles Wilkins at the University of California-Riverside. They hope to tap several collections of high quality digital spectra measured in various analytical laboratories for internal use. This team has put together an automated algorithm for evaluating the spectra of this collection [8].

2.2.6 Infrared Data Committee of Japan (IRDC) This organization has distributed IR spectra in printed form on edge-punched cards since 1961 [1d]. About 19,000 cards are now available. In 1980–85, the peak wavenumbers and intensities were extracted and entered into a computer file. Search software for the database has been prepared. A search involves entering wavenumbers and intensities in order of decreasing intensity; no-band regions can be specified. Spectra which are retrieved in a search are listed in order of the probability of being a correct match. The publisher of the IRDC cards is also marketing the above system in magnetic tape form. The possibility of fully digitizing the IRDC spectra has been discussed, but no decisions have been made.

2.2.7 Collection of National Chemical Laboratory for Industry, Japan The integrated online "Spectrum Database System" [6], which includes collections of NMR, ESR, IR, Raman, and mass spectra, also makes available a database of 22,500 infrared spectra. All spectra were determined at the NCLI under carefully controlled conditions. Data were transferred in digital form directly from the FT-IR instrument on which they were determined to the database. The database is available online to users in Japan.

2.2.8 American Society for Testing and Materials Collection Comprehensive indices coded by ASTM Committee E-13.03 for the infrared spectra from most of the older general collections are available from Chemir Labs, Sadtler Research Labs, and on-line on the Canadian Scientific Numeric Data System. Data for 145,000 compounds are included.

2.2.9 Other Collections Many hard-copy collections of IR spectra exist. For a comprehensive list of the numerous older collections, the reader is referred to the bibliography given in *The Coblentz Society Desk Book of Infrared Spectra* [9]. Paragraphs 2.2.3 and 2.2.5 describe new collection efforts aimed at the production of computerized IR databases. In addition, there are apparently several similar efforts now being initiated in Europe, notably at the University of Essen [10].

2.3 Mass Spectra

2.3.1 The Wiley Registry of Mass Spectral Data This collection has been put together and is maintained by F. W. McLafferty at Cornell University. The database, available from John Wiley & Sons, Inc. on magnetic tape or in a CD-ROM version, contains 123,704 spectra of 108,173 compounds evaluated using a Quality Index algorithm [11] (see discussion below). Replicate spectra of a given compound are included. The magnetic tape version is distributed without search software, although software for matching unknown spectra which is tailored to this database is available free of charge from Cornell University [12–17].

2.3.2 The NIST/EPA/MSDC Mass Spectral Database This database was originally put together by Drs. S. R. Heller and G. W. A. Milne of EPA and NIH, and called the EPA/NIH Mass Spectral Database. Since 1978, this database has been jointly administered by NIST and EPA, and new spectra are identified in the published literature, collected in complete form from the original authors, and evaluated by the Mass Spectrometry Data Center (MSDC), Nottingham, England. The current database consists of 43,005 spectra, each one corresponding to a unique chemical compound. Spectra in the current version of the database were selected from an archive of 79,000 mass spectra and evaluated using a Quality Index algorithm, based on—but not exactly the same as—the algorithm developed by F. W. McLafferty to evaluate the Wiley database [18,19]. (The Quality Index evaluations are discussed in detail in sec. 4.) The database is distributed on tape without search software, and in a PC version with search software and elementary matching software. A new update, which will include several thousand new spectra, is being prepared for release in the fall of 1988. The corresponding PC-version will incorporate structural information on all compounds in the database, as well as several new modes of matching spectra of unknown compounds to spectra in the database.

2.3.3 The Merged Wiley/NBS Registry of Mass Spectral Data The Wiley and NBS/EPA/MSDC collections are also available from John Wiley & Sons in a merged version, which has a total of 130,544 spectra (number of duplicate spectra in the two databases, 36,847). The merged database is available on tape and CD-ROM. A book version of the Merged Database is being published [20].

2.3.4 The Eight Peak Index The primary publication of the Mass Spectrometry Data Center, (Royal Society of Chemistry, Nottingham, England) is made up of a set of seven volumes [21] including 65,000 eight-peak spectra of 52,332 compounds indexed by molecular weight, chemical formula, and most abundant ions. This collection of partial spectra is also available on tape. The collection includes many of the same spectra included in the Wiley and NBS/EPA/MSDC collections. All of these collections of mass spectra have been put together incorporating older (non-computerized) data collections such as the spectra from the API Project 44 [1a], the Thermodynamics Research Center [1b], and the American Society for Testing and Materials (ASTM) [1e].

2.3.5 Collection of National Chemical Laboratory for Industry, Japan The integrated online "Spectrum Database System" [6], which also includes collections of NMR, ESR, IR, and Raman spectra has a database of 10,000 mass spectra which were determined in the NCLI laboratories as part of the larger project. The system was recently made available to users in Japan.

2.3.6 Other Collections (a) Japan Information System for Science and Technology (JICST) has an online "Mass Spectral Database System" searchable by name, formula, Chemical Abstracts Registry Number, and peaks. This system uses the NIST/EPA/MSDC database augmented by a collection of 6,000 spectra from the Mass Spectrometry Society of Japan. (b) Dr. D. Henneberg (Max-Planck-Institut fur Kohlenforschung) has a collection of approximately 12,000 spectra, to which he is adding with the intention of building a database [22].

3. Methods of Building Spectral Collections

While the above lists make it clear that many collections of spectra for use in analytical chemistry laboratories are available, it is also evident that Thomas Isenhour's complaint [3] that none of the collections contain more than about 100,000 spectra is also substantially correct. In order to understand why the sizes of available collections are so small even after several decades of collection effort (even excluding infrared spectroscopy, where earlier collections became less useful with the advent of new instrumentation), it is of interest to examine the techniques which are commonly used to collect spectra for such databases. This discussion will also consider how the nature and quality of a database is influenced by the way in which it has been put together.

3.1 Laboratory Efforts

The analytical chemistry databases listed above include several examples of collections which have been put together in a single laboratory by systematically determining spectra of large numbers of chemical compounds for the specific purpose of building a database. The high quality collections of infrared spectra of compounds from the Aldrich and Sigma catalogues put together by Nicolet, an instrument manufacturer, are an example of this approach.

Another example is the integrated database system put together by the National Chemical Laboratory for Industry (Japan), which includes mass spectra, IR spectra, ^{1}H and ^{13}C NMR spectra, as well as ESR and Raman spectra, all determined in the NCLI laboratories under carefully controlled conditions [6]. In addition to providing an excellent example of a carefully constructed collection of spectra, this system also is perhaps the most fully realized example of a trend which will undoubtedly become important in the future—the use of integrated databases incorporating more than one kind of spectrum.

Databases put together under this strategy are generally of high quality, since the purity of the compounds used as well as the instrument parameters can be controlled by the party building the database. In the case of the integrated database, there is the further advantage that the correctness of the data can be cross-checked by examining complementary information obtained from different techniques.

In spite of the obvious advantages of this approach, however, it must be admitted that this type of database-building effort is expensive and relatively slow. The NCLI effort, for example, has required support for a laboratory effort including IR, mass spectral, and NMR instrumentation during approximately the past dozen years; the overall database index now contains 17,000 compounds [6]. The Thermodynamic Research Center at Texas A&M University sponsors a collection effort through laboratory measurements which generates about 75 spectra per year; again, the quality of the spectra is excellent, but one could never hope to build a large database by adding spectra at this rate.

3.2 Collections Put Together through Donations of Spectra from Diverse Laboratories

Many of the collections listed above have been put together by soliciting donations of spectra from many different laboratories. The Coblentz Society collection of IR spectra [7] and the American Petroleum Institute (API) Project 44 [1a] collections of several kinds of spectra are examples of successful efforts of this nature. This approach has the obvious advantage that when a cooperative pool of donors exists, a database can be built relatively quickly and inexpensively.

On the other hand, when spectra are obtained from many different laboratories, there will inevitably be large variations in the quality of the data, not to mention differences in spectra due to the use of instruments of varying design. For example, the mass spectral collections include spectra from both magnetic sector and quadrupole instruments, which may have different types of mass discrimination, and therefore may give slightly different spectra for the same compound. However, the main problem associated with this collection technique is that completion of a collection project necessarily depends on the labor of volunteers. In general, the most successful efforts have been made when the management of a laboratory made the database collection a high priority work item (such as the petroleum industry's generation of the API Project 44 Collection). When the effort is purely voluntary—something which is done only when other (high priority) work assignments have been completed—experience has demonstrated that the time-consuming task of preparing data for transfer to a collection is rarely actually undertaken.

3.3 Collection of Data from the Literature

A large number of scientific databases are composed by abstracting data from the scientific literature. This approach can also be applied to the construction of a spectral database for analytical use, thus obviating the need for achieving cooperation from donors of spectra. The most successful example of this type of database is the Wiley Registry of Mass Spectra, put together by F. W. McLafferty at Cornell University. As a result of the incorporation of spectra from the open literature, the database has grown dramatically in recent years, achieving as noted above a size of 123,704 spectra, up by about 50,000 over a period of some four or five years.

The database one obtains using this strategy, however, has a somewhat different nature from the databases built up through dedicated laboratory measurements or donations of spectra directly from the laboratories in which they were measured. The spectral data reported in scientific papers are often incomplete, either because the journals do not have sufficient space to publish entire spectra, or because the determination of a spectrum was not the primary motivation of the work reported in the literature. Therefore, a database built up with a large component of spectra from the scientific literature will include mainly partial spectra. The mean size of a mass spectrum in the Wiley Registry is 29 peaks, which can be compared with the mean size of the spectra in the NIST/EPA/MSDC Mass Spectral Database, 60 peaks (i.e., the mean size of the spectra taken from the literature is 13 peaks/spectrum).

4. Automated Evaluation of Spectral Collections

Spectral collections which are put together by laboratories which determine each individual spectrum are evaluated as they are built, and should not require much additional evaluation. However, when spectra come from a variety of sources, through donation schemes or literature acquisition, it is important to determine the quality of the spectra, and when a collection contains more than a few thousand spectra, it is obviously advantageous to have schemes whereby the spectral quality can be examined in some automated fashion. Such an approach to the evaluation of infrared spectra has recently been reported; a sheme was developed especially for use with the University of California-Riverside "Clearinghouse for Digital Infrared Spectra" [8]. Since this scheme is new, however, few details are available about its successes and/or failures when used with an actual database.

An automated evaluation scheme for mass spectra has been in use for many years, and the successful use of automated algorithms, as well as the kinds of problems which have been encountered, can be documented. The so-called Quality Index algorithm for mass spectra was originally proposed in 1978 by Speck, Venkataraghavan, and McLafferty [11], who put together an automated examination of various factors a trained mass spectrometrist would use in evaluating spectral quality. These included: (1) energy of the ionizing electrons; (2) presence of peaks at masses higher than the molecular weight of the compound; (3) presence of "illogical" peaks, which would not normally be formed in a compound of a particular formula; (4) whether or not relative isotopic abundances were correctly represented in the spectrum; (5) the total number of peaks in the spectrum (a measure of the completeness of the spectrum); (6) the mass of the lowest peak reported in the spectrum (another measure of completeness); and (7) the source of the spectrum.

Each factor was associated with a simple equation designed to give a numerical grade ranging from 0 to 1. For example, the so-called Quality Factor for the low mass limit was assigned by examining the mass of the lowest peak reported in the spectrum (M_{min}) and comparing it to the molecular weight of the compound (MW), using the equation:

$$QF=(MW-M_{min})/(MW-39)$$

(and QF was taken to be 1.0 for all compounds with molecular weight lower than 40). The final Quality Index (QI) for the spectrum was arrived at by multiplication:

$$QI=QF_1 \cdot QF_2 \cdot QF_3 \cdot QF_4 \cdot QF_5 \cdot QF_6 \cdot QF_7 \cdot (1000).$$

Note that since the various factors are multiplied (rather than added) to achieve the final grade for a spectrum, a value of zero or a very low value for any single factor will lead to a low value for the spectrum as a whole. Furthermore, a spectrum receiving a rather high grade, but a grade less than unity, for each of the seven factors will end up with a low Quality Index value; $(0.95)^7 \times 1000 = 698$.

The same approach was used by scientists putting together the NIST/EPA/MSDC Mass

Spectral Database [18,19], who omitted the seventh Quality Factor listed above (the source of the spectrum), and added some additional factors, namely: (1) stated sample purity; (2) whether or not the mass spectrometer has been calibrated for the measurement, and, if it has, the availability of the calibration data; (3) the presence of a peak at mass 28 (taken as evidence for the presence of air); (4) evidence for detector saturation; and (5) if the spectrum does not contain a peak having a mass equal to the molecular weight, the highest mass peak which is included (again, an indicator of the completeness of the spectrum).

In addition, many of the algorithms originally formulated by the Cornell team were modified for use with the NIH/EPA database. The modifications were based largely on analyses of the statistics of the individual quality factor values obtained for the spectra in the database. That is, it was assumed that (a) the standard deviation of the values obtained for any Quality Factor should be roughly proportional to the spectral significance of the property being measured; (b) the mean value of any given Quality Factor calculated for all the spectra in the database, should be 0.9 or greater, and (c) a Quality Factor should have a value of zero only in extreme cases. This is another way of saying that any Quality Factor which penalizes essentially all spectra in the database, or very few spectra, is not giving us any useful information for distinguishing between poor and good quality spectra. Thus, the modifications generally involved changing the equations to make the penalty greater or smaller, depending on the statistics observed. For example, the "low mass limit" Quality Factor given above was found to be weighted too strongly, and was modified to:

$$QF=[(MW-M_{min})/(MW-29)]^{1/2} \text{ for } MW<179,$$

$$\text{and } QF=[(MW+179-2M_{min})/(MW+179-58)]^{1/2} \text{ for } MW>179.$$

In the NIST/EPA/MSDC database, until recently the protocol for putting together the database from the larger archive of spectra involved (1) calculating the Quality Index (QI) value for each spectrum in the system; (2) when there was more than one spectrum of a given compound, selecting from among those spectra by taking the one with the highest QI value for inclusion in the database. The spectra were not at any time examined visually by a mass spectrometrist; all judgements and selections were made using the automated procedure.

In general, the calculation is very effective in choosing between good spectra and poor spectra. However, in the 1986 edition of the database, it was noted that there were instances in which the algorithm led to the selection of a poor spectrum over several good spectra. In other cases, good spectra were found which had been assigned very low Quality Index values.

An analysis was made to identify the factors contributing to the observed problems. It was found, for instance, that spectra legitimately containing a large peak at m/z 28 were receiving low ratings because of the identification of that peak with the presence of air; the algorithm was modified to require the simultaneous presence of m/z 28 and m/z 32 with a ratio approximately the same as that one would observe for an air sample. Some of the fragmentation processes considered by the algorithm to be "illogical" were found to be important for certain types of compounds; as a result, all of the spectra of these compounds were receiving very low Quality Index values. For example, the "illogical loss" algorithm penalized all spectra in which there was an ion 2 mass-units below the parent molecular ion, that is, in which there was a fragmentation process consisting of a loss of H_2 (or 2 H-atoms) from the molecular ion. This dissociation is very important for low molecular weight alkanes, and all alkane spectra were heavily penalized. The most abundant ion in the mass spectrum of ethane is at m/z 28 ($C_2H_4^+$), and results from an "illogical loss" of two mass units, and therefore all spectra of ethane had Quality Index values of zero.

Appropriate modifications to the algorithms were carried out, and the database was regenerated. The archive contains some 16,000 spectra which are replicates; the new calculation resulted in the replacement of 620 spectra by other spectra from the archive. A visual examination of these 620 pairs revealed that 50% of the changes had resulted in the selection of a spectrum of lower quality than that originally included in the database. Some of these replacement pairs are shown in figures 1–3.

Figure 1 shows two mass spectra of HBr. At the last revision of the Quality Index calculation, the spectrum on the top replaced the spectrum on the bottom which contains HCl impurity peaks and so much water that m/z 18 is the major peak. Note that although the current algorithm results in the choice of the better spectrum, the difference in the

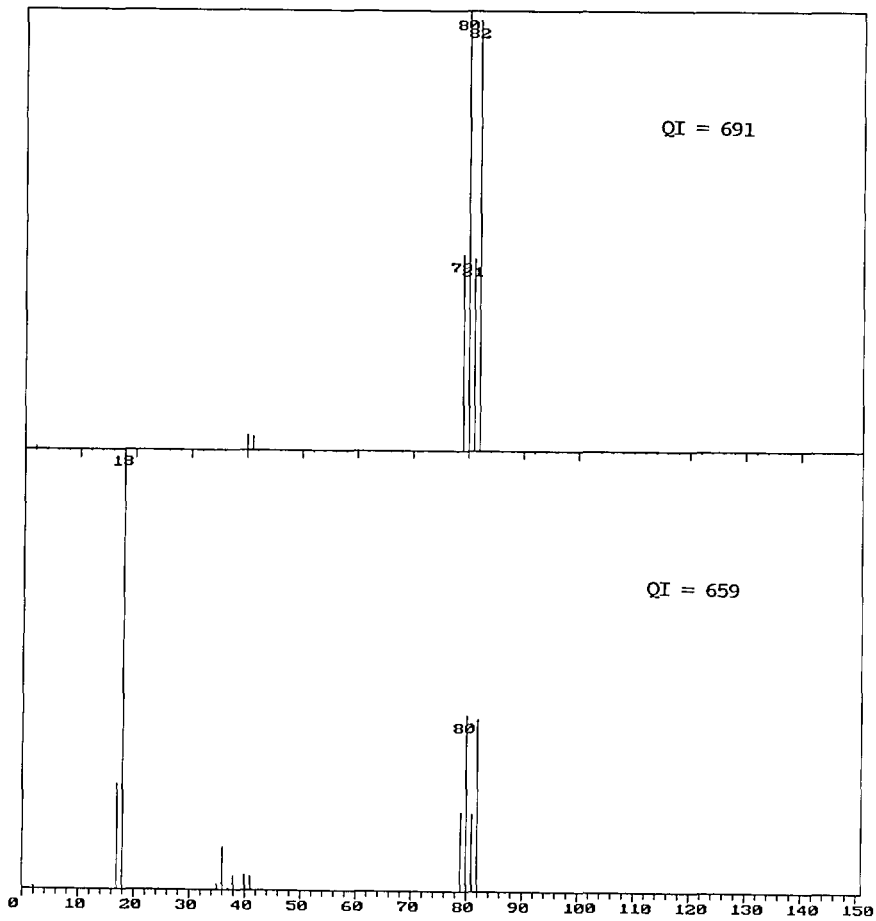


Figure 1. The mass spectrum of HBr shown on the bottom, containing water as the major component, was replaced by the spectrum shown on the top when the Quality Index calculation was revised (see discussion in text).

QI values between the good spectrum and the very bad spectrum is only 32 points

Figure 2 shows four spectra of thiourea. The spectrum on the top (A) is missing a major peak at *m/z* 43 (it appears that this peak has been misidentified as *m/z* 42), and an extra impurity peak at *m/z* 44 (or 45). That incorrect spectrum was formerly selected for the database; the "illogical fragmentation" algorithm did not recognize the incorrectly identified peaks. The spectrum (A) was replaced by the revised *QI* calculation with the spectrum (B) shown second, which now has a *QI* value 18 points higher than that of (A). Although spectrum (B) appears to be somewhat more complete than spectra (C) and (D), it clearly suffers from detector saturation, and therefore would be considered by an expert to be inferior in quality to both spectra (C) and (D). Curiously, the bad spectrum (A) receives the same *QI* grade as the good spectrum (C). Since the fragmentation of this parent ion does lead to the formation of an ion of *m/z* 42, it is unlikely that any algorithm could have detected the mistake in spectrum (A).

Figure 3 shows two spectra with Quality Index values which are within two points of one another. The spectrum with the higher *QI* value contains peaks, for example, at masses 41 and 44, which can only originate from an impurity.

An examination of these examples leads to the conclusion that *this type of Quality Index algorithm could not have done any better at selecting the best spectrum from among replicates*. With more fine tuning, this algorithm as it is presently constituted will never do any better. In setting up an evaluation-selection system of highly arbitrary equations, one is implicitly accepting that some statistical fraction of the spectra selected will be spectra which are not the best examples available in the archive. For instance, the recently-introduced Quality Factor, designed to penalize detector satu-

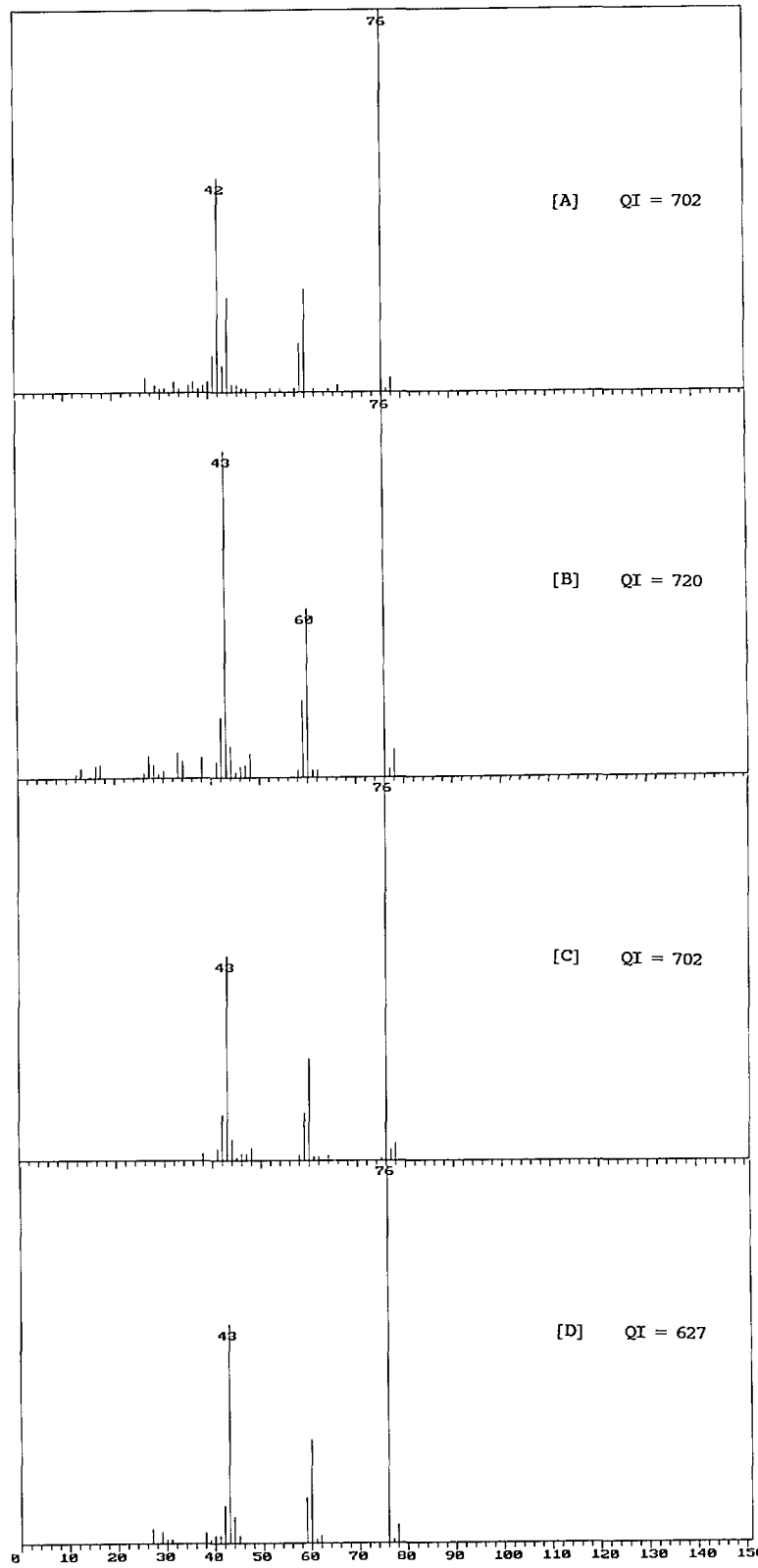


Figure 2. Four mass spectra of thiourea. In spectrum (A), m/z 43 has been misidentified as m/z 42; this is the spectrum originally selected by the *QI* calculation. Revision of the algorithm resulted in the choice of spectrum (B), which exhibits detector saturation. Spectra (C) and (D) (not selected by the program) are better quality spectra than (A) and (B).

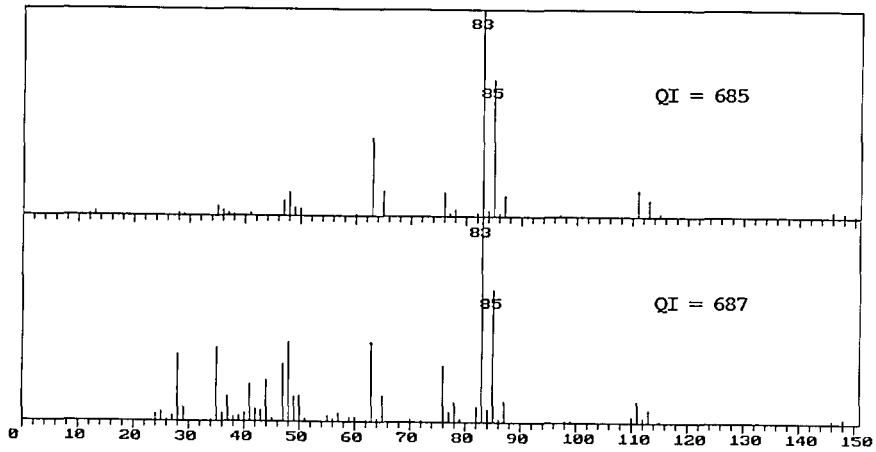


Figure 3. Two mass spectra of dichloroacetyl chloride exhibiting Quality Index values which differ by only 2 points. The lower spectrum, which has the higher *QI* value, contains peaks, for example at masses 41 and 44, which can only originate from an impurity.

ration, does so by searching for spectra having one or more additional peaks similar in magnitude to the base peak (peak of maximum abundance in a mass spectrum). Of course, some spectra legitimately have peaks of such magnitude, and they will be penalized; other spectra may be significantly saturated, but still pass such a test. The authors discuss this problem and conclude that *these errors can be tolerated* if the algorithm catches a large fraction of saturated spectra.

Until a truly "expert system" approach to the evaluation of analytical mass spectra is devised, it appears that the only possible procedure for selecting only the best available spectrum of each compound from an archive is to (1) use the existing Quality Index calculation as a rough first selection procedure, and (2) have an expert carry out a visual selection from among those replicate sets for which the Quality Index values are within 200–300 points of one another. This is the procedure now being carried out on the NIST/EPA/MSDC Mass Spectral Database, preparatory to release of the next update.

5. Acknowledgment

Information on NMR databases was collected by Dr. Bruce Coxon, whose help is gratefully acknowledged.

6. References

[1] Examples of early collection efforts include: (a) API Research Project 44, Carnegie Institute of Technology, Pittsburgh, Pennsylvania; a subscription service producing data sheets for mass spectra, NMR, and other spectra beginning in the late 1940s; (b) Thermodynamics Research Center Spectra, Texas A&M University; the continuation of the older API Project 44; (c) Sadtler Spectral Data Sheets Including IR, Raman, ^{1}H and ^{13}C NMR, Sadtler Research Laboratories, Spring Garden St., Philadelphia, Pennsylvania; for a history of this effort, see: Sadtler, P., and Sadtler, T., History of 'Sadtler' Spectroscopy, Applied Spectrosc. **39**, xix (1985); (d) IRDC Infrared Spectral Cards, Japan Infrared Data Committee, c/o Prof. I. Suzuki, The University of Tsukuba, Tsukuba Scientific City, Ibaraki, Japan; (e) Spectral collections of the American Society for Testing and Materials (ASTM), Philadelphia, Pennsylvania.

[2] The nature of the survey presented in this paper is such that it is necessary to mention comercial databases by name. In no instance does such identification imply endorsement by NIST.

[3] Isenhour, T. L., Spectroscopic Data Bases, J. Chem. Inf. Comput. Sci. **26**, 2A (1986).

[4] Chemical Information System: An online collection of databases administered by Fein-Marquart Associates, Inc., 7215 York Road, Baltimore, Maryland 21212.

[5] Scientific and Technical Information Network (STN International): The online datbase administered by the Chemical Abstracts Service.

[6] Yamamoto, O., Someno, K., Wasada, N., Hiraishi, J., Hayamizu, K., Tanabe, K., Tamura, T., and Yanagisawa, M., An Integrated Spectral Data Base System Including IR, MS<^{1}H-NMR, ^{13}C-NMR, ESR and Raman Spectra, Anal. Sci. **4**, 233 (1988).

[7] Coblentz Society Spectra, P.O. Box 9952, Kirkwood, Missouri 63122.

[8] Griffiths, P. R., and Wilkins, C. L., Quality Criteria for Digital Infrared Reference Spectra, Appl. Spectrosc. **42**, 538 (1988).

[9] The Coblentz Society Desk Book of Infrared Spectra, (C. Craver, editor), The Coblentz Society, Inc., Kirkwood, Missouri (1977).
[10] Griffiths, P., personal communication.
[11] Speck, D. D., Venkataraghavan, R., and McLafferty, F. W., A Quality Index for Reference Mass Spectra, Org. Mass Spectrom. **13**, 209 (1978).
[12] McLafferty, F. W., Hertel, R. H., and Villwock, R. D., Org. Mass Spectrom. **9**, 690 (1974).
[13] Pesyna, G. M., Venkataraghavan, R., Dayringer, H. E., and McLafferty, F. W., Anal. Chem. **48**, 1362 (1976).
[14] Mun, I. K., Venkataraghavan, R., and McLafferty, F. W., Anal. Chem. **49**, 1723 (1977).
[15] Atwater (Fell), B. L., Venkataraghavan, R., and McLafferty, F. W., Anal. Chem. **51**, 1945 (1979).
[16] Stauffer, D. B., Sharaf, M. A., Dromey, R. G., Guo, C. J., and McLafferty, F. W., The 31st Annual Conference on Mass Spectrometry and Allied Topics, Boston, Massachusetts, May 8–13, 1983, p. 558.
[17] McLafferty, F. W., Cheng, S., Dully, K. M., Guo, C. J., Mun, I. K., Peterson, D. W., Russo, S. O., Salvucci, D. A., Serum, J. W., Staedeli, W., and Stauffer, D. B., Int. J. Mass Spectrom. Ion Phys. **47**, 317 (1983).
[18] Milne, G. W. A., Budde, W. L., Heller, S. R., Martinsen, D. P., and Oldham, R. G., Org. Mass Spectrom. **17**, 547 (1982).
[19] Terwilliger, D. T., Behbehani, A. L., Ireland, J. C., and Budde, W. L., The Status and Evaluation of a Mass Spectral Data Base, Biomed. Environ. Mass Spectrom. **14**, 263 (1987).
[20] McLafferty, F. W., and Stauffer, D. B., Wiley/NBS Registry of Mass Spectral Data, John Wiley & Sons, in press.
[21] Eight Peak Index of Mass Spectra, The Royal Society of Chemistry, Distribution Centre, Blackhorse Road, Letchworth, Herts. GS6 1HN, England.
[22] Henneberg, D., personal communication.

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

The Structural Ceramics Database: Technical Foundations

Volume 94 Number 1 January–February 1989

R. G. Munro, F. Y. Hwang[1], and C. R. Hubbard[2]

Ceramics Division
National Institute of Standards and Technology
Gaithersburg, MD 20899

The development of a computerized database on advanced structural ceramics can play a critical role in fostering the widespread use of ceramics in industry and in advanced technologies. A computerized database may be the most effective means of accelerating technology development by enabling new materials to be incorporated into designs far more rapidly than would have been possible with traditional information transfer processes. Faster, more efficient access to critical data is the basis for creating this technological advantage. Further, a computerized database provides the means for a more consistent treatment of data, greater quality control and product reliability, and improved continuity of research and development programs.

A preliminary system has been completed as phase one of an ongoing program to establish the Structural Ceramics Database system. The system is designed to be used on personal computers. Developed in a modular design, the preliminary system is focused on the thermal properties of monolithic ceramics. The initial modules consist of materials specification, thermal expansion, thermal conductivity, thermal diffusivity, specific heat, thermal shock resistance, and a bibliography of data references. Query and output programs also have been developed for use with these modules. The latter program elements, along with the database modules, will be subjected to several stages of testing and refinement in the second phase of this effort. The goal of the refinement process will be the establishment of this system as a user-friendly prototype.

Three primary considerations provide the guidelines to the system's development: (1) The user's needs; (2) The nature of materials properties; and (3) The requirements of the programming language. The present report discusses the manner and rationale by which each of these considerations leads to specific features in the design of the system.

Key words: ceramics; computerized database; material properties; Structural Ceramics Database; user-friendly.

Introduction

Technical advances in materials research are occurring at a rapid pace in all aspects of the development and refinement of advanced ceramics. As a result, technical data are proliferating at an exponentially increasing rate. Industries may reasonably anticipate new technological opportunities to accompany these advances in materials research. However, some of these opportunities will be lost if the new data are not successfully communicated to the design engineers who can transform the data into new or better products. The traditional communication routes for disseminating technical data frequently have slow diffusion rates

[1] Guest Scientist from: Material Research Laboratories, Taiwan, Republic of China.
[2] Current address: Oak Ridge National Laboratory, Oak Ridge, TN 37831.

and may be cumbersome to use, especially when disciplinary boundaries are crossed. To capture the data and to fashion it into an expeditiously useful form, advances in computerized information systems are being pursued vigorously in a wide range of scientific and engineering fields [1-7], and involve national agencies [8], professional societies [9], and dedicated nonprofit organizations [10]. The present report discusses the initial results of the effort at the National Institute of Standards and Technology to develop a new database system that will be focused on critical material properties of advanced structural ceramics. The current effort is being conducted in conjunction with the research program at the Center for Advanced Materials established at the Pennsylvania State University by the Gas Research Institute.

The rapidly increasing importance of computerized information systems is readily apparent to even the most casual observer. The number of information sources in technical areas is enormous. The subject matters within these areas are increasingly diversified. There are a steadily proliferating number of specializations within each subject area. Each specialization generates technical terminology not in common with other specializations. And, the objectives for obtaining, developing, or using the data are unlimited.

Coping with this abundance and diversity of information is the function of the computerized database. By using the logistical power of computers, data can be stored, sorted, searched, retrieved, and used in a very small fraction of the time required by manual methods. Further, data stored in a computerized database may form the basis of a technical "corporate memory" because the availability and usefulness of the data persist beyond the lifetime of the project that generated the data. As a result, there is not only a more rapid utilization of technical advances, but also a reduction of wasteful duplication of efforts. This powerful processing of information can create improved perceptions of research strengths and weaknesses and, hence, may provide improved managerial vision for future research planning.

The Structural Ceramics Database (SCD) system is being developed so that these capabilities can be used to help industry in taking advantage of newly emerging specialized ceramics. This objective has two inherent requirements. The system must include critical data, and the system must be easy to use. The former requirement pertains to the content of the database. The latter requirement pertains to how the computerized system is constructed. The SCD project, therefore, has both a data acquisition component and a system development activity.

The first phase of the SCD project was focused on the development of a preliminary software system. The initial emphasis, therefore, was on technical considerations, and the result was a preliminary system that has fully functioning storage, search, and retrieval capabilities.

The first step of the project was to consider the needs of the user. The user's requirements formed the basis for specifying the technical requirements of the underlying database management system (DBMS) that would be selected as the programming language. To further guide the development of the system, a specific application area, high temperature gas-fueled heat exchangers, was selected, and the gathering of data was started. These factors were combined in determining the initial structure of the database system that will be discussed in the following sections.

Issues

The development of the SCD system requires the resolution of three sets of issues: the design issues that result from consideration of the user's needs [11-12], the technical issues that result from the constraints of the programming language [13-14], and the materials properties issues that result from the nature of the materials and the particular properties that are to be included in the database.

The User Interface

First and foremost, the system must be easy to use. The system should place very little demand upon the user in terms of technical knowledge of computers or computer programming. Once started, the system should guide the user at each step of the interactive session, always clearly indicating the user's options. Never should the user need to refer to an operations manual.

At the same time, the intelligence of the user must be respected. The user of a database wants to extract information that usually has a well defined scope and content. The system must make it easy for the user to pose questions to the database in a clear and concise manner, and the answers to the questions must be presented to the user quickly and in a readily understood format. Those answers should also be expressed in units with appropriate

dimensions and terminology that are convenient and comfortable for the user. The answers must also be comprehensive in the sense that the user's ranges of conditions and questions are anticipated, so that the questions *can* be asked, and that there are data available to answer them. Further, whenever the user wants additional information regarding any specific data, the system should provide references to readily available literature.

To accomplish these features, the design of the database should include three primary characteristics: simple screens; menus with light-bars; and online help.

In many respects, what appears on the screen is the immediate link between the database and the user. The appearance of the screen, therefore, should not assault the user's senses but, rather, should focus the user's attention on the central concern. Too much information presented to the user on the screen creates confusion and distraction. To make a screen simple, the program sometimes must bear the responsibility for considerable preparatory steps prior to presenting the principal screen. As a result, the design of a simple screen is often, ironically, more difficult than the design of a complicated screen.

The use of a menu is probably the most important means of simplifying the interaction between the database and the user. When menus are used, there is no need or requirement for the user to remember program commands that are too frequently crytic, obscure, or unclear. The options are always presented to the user who only needs to select one of them. As a result, there are fewer entry errors. Further, the database system is always in control of the program flow, i.e., the user can only ask the database to do operations that it is ready to do. If primitive DBMS commands are accessed directly, it is possible to ask a program to execute operations before all of the preparatory steps have been completed. In this situation, the program can "hang" or "crash". With a menu-driven system, the responsibility for maintaining the proper program flow rests with the design of the system.

Menu systems are further enhanced by the use of light-bars, a rectangular area that appears on the screen as a highlighted region. The light-bar highlights one option and can be moved by the user to any other option. To make a selection from among the options, the user merely moves the light-bar to highlight the desired option and presses the return key. Exactly the same set of keystrokes are used with every such menu. Hence, the user's need to be familiar with the mechanical aspects of the keyboard are minimized. Light-bars can also be programmed so that a pointing device such as a mouse can be used, thereby entirely eliminating keystrokes for menu selections. Also important is the fact that the user *sees* what the choice is when the light-bar highlights an option.

The information and the options presented to the user of a materials property database must relate to the technical content of the database. Consequently, the words and terminology used on the screen must be tailored to the technical material. Technical terminology is rarely standardized across all disciplines. Therefore, it is essential that online help be available to the user at any time to explain the options and the terminology, and to provide references to the literature where an in-depth discussion on the subject matter can be found. Online help provides information to the user immediately, when it means the most to the user.

Materials Property Issues

The design of a database necessarily requires consideration of the information to be contained in the database. Data in a database occurs in four types: numeric (numbers), character (words), logical (true or false), and date (month, day, and year). Before the database can be prescribed, the type of data that it is to contain must be known. The specification of what specific information is needed for a materials property database is less obvious than it might seem at first thought. For example, consider what is required to identify a particular structural ceramic material. What characteristics uniquely define the material? Conversely, can the material be grouped with other materials as part of a more general class?

It is widely recognized that the name of a ceramic material does not provide an adequate description of the material. For example, there are many forms of silicon nitride. All silicon nitrides have the same primary chemical formula, Si_3N_4, but the sintering aids, impurity components, porosity, and microstructure are different. This question of identifying a ceramic material is being investigated currently by the ASTM Committee E-49, Computerization of Material Property Data. Their deliberations indicate that at least 10 categories of supporting information, also called metadata, table 1, are desirable for the specification of an advanced material [15]. These descriptors only identify the material and do not provide any of the material properties.

Table 1. Preliminary guidelines from ASTM Committee E-49 regarding the categories of information needed to describe a material for database purposes

Material Class
Specific Material within a Class
Material Designation
Material Condition
Material Specification
Producer or Source of Material
Producer Lot Number and/or Assigned Reference Number
Product Form
Material Composition
Fabrication History

Each material property to be included in the database also must be given carefully selected specification data. What is the property? What method was used to measure it? Are the data of reasonable quality?

The range of properties that should be included in the database depends on the application or the intended use of the database. A general database that is not focused on any particular application would need to include far too many properties than would be practical. Rather than construct a database that is entirely comprehensive, it is preferable that the database be developed for a well focused application for which the critical data can be clearly determined.

For any given property, an important issue pertaining to the usefulness of the data is how much supporting detail should be available to describe how the property value was determined. What depth of information is required concerning the method used, the conditions of the experiments, or the statistical treatments that may have been applied to the data? These questions are encountered whenever data are reported in the technical literature. Technical papers are usually required to provide in some manner a complete description of all experimental apparatus, experimental procedures, and data analysis. A database is not intended, and should not be expected, to replace a technical paper. However, it may be important to know what experimental techniques were used to evaluate the property. Different methods may subject the material to different conditions, and hence, the results from one technique may be more appropriate to the user's application than other techniques. To accommodate the need to be clear, succinct, and complete with respect to property data and its determination, a bibliography of technical references should be maintained as part of the metadata used to describe and record material properties.

The final concern regarding the data is the quality of the values recorded in the database. Some assessment of the quality of the data, its accuracy or reliability must be provided with the value that is recorded in the database. If there are limitations on the validity of the value, the user needs to be forewarned of the limitations. The design of the database, therefore, should include metadata fields in which the quality and limitations of the data can be noted.

Constraints of the Programming Environment

The programming environment for the development of a database is most conveniently and wisely taken to be a commercially available database management system (DBMS). A DBMS is essentially a language that can be used to tell the computer how to store and retrieve data. Commercial DBMS packages contain many highly refined features that greatly facilitate the creation of a database architecture and provide the essential means to search for information stored within the resulting structure. These features, while sophisticated and desirable, may also be viewed as constraints on the database design. Thus, it is important to identify the technical features that are necessary to secure compatibility with the data that are to be included in the database and to ensure the fulfillment of the requirements of the user.

The SCD must consist of many material properties and characteristics, including the materials specification, chemical composition, microstructure, mechanical properties, and thermal characteristics. Each piece of information that is to be included in the database must be alloted a distinct amount of space where the information can be stored. In the simplest structure of a database, the pieces, called fields, are concatenated to form a single collection of information, called a record,

Field 1	Field 2	Field 3	Field 4	Field 5	Field 6

as illustrated in the diagram. In this example, six fields of different sizes are joined together on one line to form one record. The complete database may then be visualized as a collection of several such records, one per line, with each field aligned to form a column. A collection of three records would look like the following matrix:

(Field 1)	(Field 2)	(Field 3)	(Field 4)	(Field 5)	(Field 6)
(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)

In principle, this record structure can be used for any database. In practice, however, this structure can be somewhat awkward and inconvenient.

To illustrate this point, consider a simple database that contains only selected thermal properties of materials as a function of temperature. For the purpose of discussion, assume that the fields for this hypothetical database are restricted to the name of the material, the temperature, the thermal conductivity, and the thermal shock resistance. Suppose also that results for alumina have been obtained from a published source that reported the thermal conductivity at three temperatures. Then, the database with only this data would be:

(Material Name)	(Temperature)	(Conductivity)	(Shock Resistance)
Alumina	20	29	
Alumina	500	12	
Alumina	1000	9	

The database has three records corresponding to the three temperatures at which the thermal conductivity was determined. The field containing the Material Name has the same information in each record. This redundancy of information is typical of much of the supporting information that is required to make the database useful. Many more fields with such metadata would be necessary to complete the specification of a good database on thermal conductivity, including the units of temperature and thermal conductivity, detailed information on the composition and microstructure of the material and the processing technique used to make the material, and details about the measurement method. The proliferation of such fields can rapidly lead to a high degree of redundant information in the database. In the example, the field for Shock Resistance is not only redundant, but also superfluous since no values for that property were given in the referenced report.

Redundant and superfluous data are wasteful of limited storage space, reduce the speed with which the information can be processed, and increase the potential for errors in the entry of the information into the database. To avoid this situation, a different type of database structure is needed. Instead of linking all the fields together into a single record, the fields can be divided into logical subgroups. Either a hierarchial or a relational database structure can accomplish this organization.

Hierarchial databases achieve a logical organization and space efficiency by creating a tree structure. Each new component to the database becomes a new branch in the tree. If a component is not used, that branch does not need to be created. Thus, the hierarchial system does not require any wasted space. However, maintenance or revision of the tree structure can be cumbersome, and navigating the tree from one record to another can be awkward.

Relational databases provide a more desirable structure for scientific and engineering applications that may anticipate a need for revision as the discipline progresses. To illustrate this point, consider the preceding example. The original database could be divided into three component databases, one each for materials specification, thermal conductivity, and thermal shock resistance. In a relational database system, each of these components can be maintained independently, provided that a unique relation between the components is specified and preserved. This relationship can be established, for example, by adding a new field, called a key, that identifies the source of the data in each component. In the example, the three subgroup databases could be defined as follows:

-----Subgroup----- ----------Fields----------

Materials: (ID) (Name)

0001	Alumina

Thermal Conductivity: (ID) (Temperature) (Conductivity)

0001	20	29
0001	500	12
0001	1000	9

Shock Resistance: (ID) (Temperature) (Resistance)

The greater efficiency of the relational database structure is readily apparent even in this restricted example. The addition of the ID field is sufficient to make the relationship between the subgroups

clear and unambiguous. The new structure completely eliminates the previously redundant and superfluous data. For a materials property database, a fully developed subgroup might contain as many as 50 fields. Hence, the relational database structure may be considered the preferred structure for scientific and engineering databases.

In the example, it may be noted, further, that three records were required to record the thermal conductivity at three temperatures, even though all three results were obtained from one reference. Logically, it would be easier to search and retrieve this information if all the values were contained in one record. Advanced DBMS packages provide this data structure in the form of multiple entry associated fields. In a multiple entry field, a variable number of different values can be entered. Associated fields are multiple entry fields for which there is a one-to-one correspondence between the associated entries. For example, if temperature and conductivity are associated fields, then the Thermal Conductivity subgroup described above would become:

Thermal Conductivity: (ID) (Temperature) (Conductivity)

0001	20, 500, 1000	29, 12, 9

The first entry in the temperature field, 20, corresponds to the first entry in the conductivity field, 29. Associated fields capture the entire set of relevant data in the single record, thereby making the search and retrieval operations more efficient and faster. The entire example database, recast as a relational database with associated, multiple entry fields, is reduced to only the following:

-----Subgroup----- ----------Fields----------

Materials: (ID) (Name)

0001	Alumina

Thermal Conductivity: (ID) (Temperature) (Conductivity)

0001	20, 500, 1000	29, 12, 9

Shock Resistance: (ID) (Temperature) (Resistance)

Most materials property information occurs with parametric dependencies and with many metadata fields. Hence, associated multiple entry fields in a relational database are natural and desirable refinements of the database architecture for scientific and engineering applications.

Next to the specification of the underlying database structure, the most important consideration for the programming environment is the ability to search the database for information. Indeed, it is this stability that makes databases useful and powerful. As a result, the search and retrieval features are the highlights of many of the advanced DBMS packages available commercially. In making comparisons of search and retrieval capabilities, it is important to recognize that most searches are specified in terms of the supporting metadata rather than the property itself. Many of the metadata fields are textual in character, i.e., collections of words. To find what the user has specified, the DBMS may need to search *within* a field to find the user's words embedded within the field. Not all DBMS packages permit this type of search. For materials property databases, searching within a field may be essential for characteristics such as chemical composition, microstructure, or processing conditions. A summary of the general characteristics that may reasonably be expected of a DBMS for scientific and engineering applications is given in table 2.

Table 2. Essential requirements for a database management system (DBMS) applied to scientific and engineering information systems

Relational database structure
Variable length fields
Large field lengths, in kilobyte range
Multiple entry fields
Associated fields
Large number of fields per record
Large total record sizes, in kilobyte range
Very large maximum number of records
High efficiency indexing
Concurrent indexing
Multiple field keys for indexing
Search on any field
Search within a field
Logically concatenated searches on several fields
Compatibility with other computer languages
Compiled or runtime codes

Structural Ceramics Database

The development of the SCD system has been structured in a logical sequence of steps. Initially, a particular application was selected to provide a definite focus for the choice of properties to be included in the database. The particular application

was selected after conducting a survey of the activities of the project's founding sponsor, the Gas Research Institute (GRI) [16]. GRI contractors indicated that tailoring the system for use with heat exchanger and recuperator design would be especially valuable to them in both the short-term and the long-term. Further discussions with GRI contractors and other research and industrial contacts produced a list of critical properties, i.e., those properties that are needed for the design efforts and for which current information is often inaccessible, cumbersome to find, and/or essential to the design. With a focused application and a sampling of the typical data to be included in the database, the general technical specifications were formulated. The latter specifications were coupled to the system needs, as seen from the user's point of view, to determine the overall requirements for the basic database management system (DBMS), as summarized in table 2. Based on those requirements, a commercially available DBMS was selected.

The system software design effort was then aimed at developing a fully functioning prototype system. The construction of the prototype was begun using a modular design consisting of general modules that may be used for any target application, and tailored modules that pertain only to specific material properties.

The modular design provides considerable flexibility and allows the system to be expanded or adapted to diverse applications. This design is also being exploited in a pragmatic way. It is perhaps well known that it is much easier to develop a programmer's working system than it is to develop a working *user-friendly* system. It may also be readily understood that certain data entry and review modules are necessary before data can be entered into the system and be verified. Consequently, to maximize the efficiency of the system development, phase one of this project was directed towards the development of those modules that would provide a working, preliminary system as soon as possible. The user interface modules are scheduled to be developed in the second phase of the project.

The preliminary system of the phase one effort is now complete. The preliminary system consists of modules for materials specification, thermal expansion, thermal conductivity, thermal diffusivity, specific heat, thermal shock resistance, a bibliography of data references, queries, and output. Currently, the query and output modules are rather general, for development purposes, and need to be streamlined before being used in the prototype of phase two. The other modules are ready for testing in their current forms.

Tables 3–9 summarize the contents of the current primary information modules. For some fields, such as Material Class, the entries are restricted to a small, finite, closed set. The possible entries for those fields are also shown in the tables.

Table 3. A listing of the information fields contained in the materials specification module. Where only fixed entries are allowed, the alternatives are listed below the field topic

Material class:
Monolithic ceramic
Structure class:
Polycrystalline, Single crystal, Graphitic, Amorphous
Chemical class:
Carbide, Nitride, Oxide
Chemical name
Chemical Abstract Service Number
Chemical formula
Source of the material
Manufacturer's designation for the material
Manufacturer's lot number
Product date
Standard design specification code
Organization setting standard specification
Supplementary information regarding specification
Fabrication process:
Slip casting, Tape casting, Sintering (firing), Extension, Mechanical throwing, Die pressing, Isostatic pressing, Injection molding, Hot pressing, Glass ceramics route, Glazing, Plasma spray, Chemical vapor deposition, Crushed ground
Fabrication form:
Plate, Bar, Rod, Wire, Tube, Thick film, Thin film, Powder
Fabrication history
Specimen state: Virgin, Modified
Description of specimen modification
Location of the specimen from within the fabrication form
Elemental composition of the material
Weight percents of the elemental components
Standard deviations of weight percents of elemental components
Phase composition of the material
Weight percents of the phase components
Standard deviations of weight percents of the phase components
Mean value of the bulk density of the material
Unit of density: g/cm^3, kg/m^3
Theoretic density
Grain size
Unit of grain size: μm, mm, m
Known or intended applications of this material
Supplementary notes

Table 4. A listing of the information fields contained in the thermal expansion module. Where only fixed entries are allowed, the alternatives are listed below the field topic

Measurement method:
Push-rod dilatometer, Twin tele-microscope, Interferometer,
Single crystal x-ray diffraction, Powder x-ray diffraction,
Neutron diffraction, High speed pulsed heating,
Volumetric dilatometry, ASTM C372 (Dilatometry),
ASTM E289 (Interferometry),
ASTM E831 (Thermodilatometry), Other
Sample preparation/pretreatment
Measurement notes
Quality
Cautions
Unit of temperature: °C, °F, K
Unit of thermal expansion coefficient: 10^{-6} K^{-1}
Thermal expansion coefficients at temperature:
Temperature
Bulk average value
Value along *a*-axis
Value along *b*-axis
Value along *c*-axis
Polynomial representation of principal axial coefficients $\alpha(i, i)$:
Temperature range from _____ to _____
$\alpha(1, 1)$=_____+_____ $(T/1000)$+_____ $(T/1000)^2$
$\alpha(2, 2)$=_____+_____ $(T/1000)$+_____ $(T/1000)^2$
$\alpha(3, 3)$=_____+_____ $(T/1000)$+_____ $(T/1000)^2$

Table 5. A listing of the information fields contained in the thermal conductivity module. Where only fixed entries are allowed, the alternatives are listed below the field topic

Measurement Method:
Longitudinal heat flow, Forbes' bar, Radial heat flow,
Direct electrical heating, Thermoelectric,
Thermal comparator, Periodic heat flow, Transient heat flow,
ASTM C201 (Comparative), Other
Sample Preparation/Pretreatment
Measurement Notes
Quality
Cautions
Unit of temperature: °C, °F, K
Unit of thermal conductivity: W m^{-1} K^{-1}
Temperature
Thermal conductivity

Table 6. A listing of the information fields contained in the thermal diffusivity module. Where only fixed entries are allowed, the alternatives are listed below the field topic

Measurement Method:
Center-heated long bar, End-heated long bar,
Moving heat source, Small area contact,
Thermoelectric effect, Semi-infinite plate, Radial heat flow,
High intensity arc, Flash heating, Electrically heated rod,
Angstrom's method, Modified Angstrom's method,
Temperature wave velocity,
Temperature wave amplitude-decrement, Phase lag,
Radial wave, ASTM C351 (Insulating Materials),
ASTM C714 (Thermal Pulse), Other
Sample Preparation/Pretreatment
Measurement Notes
Quality
Cautions
Unit of temperature: °C, °F, K
Unit of thermal diffusivity: m^2 s^{-1}
Temperature
Thermal Diffusivity

Table 7. A listing of the information fields contained in the specific heat module. Where only fixed entries are allowed, the alternatives are listed below the field topic

Measurement Method:
Nernst-type adiabatic vacuum calorimeter,
Modified adiabatic calorimeter, Drop ice calorimeter,
Drop isothermal water calorimeter,
Drop copper block calorimeter, Pulse heating,
Comparative method, ASTM C351 (Insulating Materials),
Other
Sample Preparation/Pretreatment
Measurement Notes
Quality
Cautions
Unit of temperature: °C, °F, K
Unit of specific heat: J kg^{-1} K^{-1}
Temperature
Specific heat

Table 8. A listing of the information fields contained in the thermal shock module. Where only fixed entries are allowed, the alternatives are listed below the field topic

Measurement Method:
Water Quench/Internal Friction, Other
Sample Preparation/Pretreatment
Measurement Notes
Quality
Cautions
Unit of temperature: °C, °F, K
Critical quench temperature difference

Table 9. A listing of the information fields contained in the bibliography module. Where only fixed entries are allowed, the alternatives are listed below the field topic

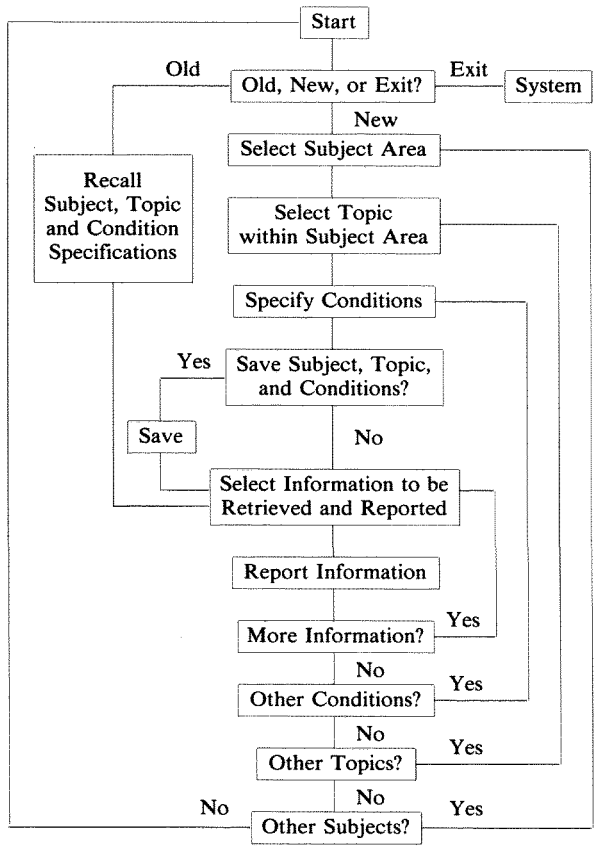
Names of authors
Total number of authors
Country of authors
Title of paper/report/article
Type of publication:
Journal, Book, General report, Contract report, Conference, Dissertation/Thesis, Patent, Magnetic tape, Private, Other
Name of publication medium
Names of editors of the publication medium
Language of publication
Chapter number
Volume number
Issue number
Page numbers
Publication date
Publisher
Sponsor
Site of conference, meeting, university granting degree, or other site
Patent number
Patent country
Initial source of abstract citation
Chemical Abstract Service abstract number
International Standard Serial Number
Number of materials discussed
Names of materials
Synonyms for materials
Chemical Abstract Registry number
Physical properties discussed:
Lattice parameter
Chemical properties discussed:
Corrosion, Corrosion products, Corrosion rate, Oxidation, Chemical reactivity, Stability
Thermal properties discussed:
Thermal expansion, Thermal conductivity, Thermal diffusivity, Specific heat, Thermal shock, Thermal emissivity
Mechanical properties discussed:
Elastic modulus, Elastic constants, Young's modulus (E), Shear modulus, Poisson's ratio, Bulk modulus, Compressibility, Strength, Tensile strength, Flexure strength, Bend strength, Fracture strength, Rupture strength, Stress-strain, Fracture energy, Fracture toughness, Toughness, R-curve, Critical stress intensity factor, Crack growth, Creep, Creep rupture, Fatigue, Cyclic fatigue

The initial materials specification module has been constructed in accordance with the guidelines evolving from the deliberations of ASTM Committee E-49, Computerization of Material Property Data. The guidelines, summarized in table 1, have been implemented by dividing the required information into 32 fields. In brief, these fields describe what the material is called, how and where it was made, and what physical and chemical characterization information has been recorded.

The materials property databases are constructed with the sets of information fields intentionally kept small. The user's first priority is to know the value of a property under a specifed condition. In general, each of the measurement methods has certain measurement conditions associated with it. Thus, identifying the method effectively identifies the conditions. If complete details of a particular measurement are needed, the user may consult the literature reference that is included in the bibliographic database. Further, a help function provides general references in which descriptions of all the related measurement techniques may be found.

Discussion

From the point of view of the user, the basic function of the database is to help the user obtain information. To do this, the database system must have a flexible, but user friendly, query capability. The essence of the query system in the SCD is illustrated in the following flow chart:



Any query session may be saved for future use. Therefore, at the beginning of a new session, the user is given the opportunity to recall a previous query or to start a new one. A new investigation begins with the selection of the central subject on which the search of the SCD will be based. The subject could be a material, a property, or a bibliographic reference. For example, to examine the properties of a silicon nitride, the subject would be ceramic materials and the topic would be silicon nitride. Further specifications, such as sintering aids, phase composition, or a maximum value for the thermal expansion coefficient, could be made in the Specify Conditions step. When the user is satisfied with the constructed query, the query conditions may be saved for future reference.

The SCD system uses the query conditions to select from the entire database only the subset of records that fulfill the user's requirements. The user may then choose to examine any part of the information in the subset. For example, having specified silicon nitride with MgO as a sintering aid and a linear thermal expansion coefficient not greater than 3.5×10^{-6} K^{-1}, the user could readily determine the variation of the thermal shock resistance with respect to fabrication process. The latter information would be obtained by specifying that the thermal shock resistance and the fabrication process be included in the list of items reported to the user.

The SCD query program is more powerful than the simplified flow chart reveals. Indeed, the program currently is too powerful to be user friendly. Within the specifications of conditions, it is possible to conduct several independent queries and then combine them into a single complex query. However, complex queries require the user to have a considerable knowledge about the structure and content of the fields in the various databases. An objective of the user interface module will be to harness the power of the query program so that the user can find precisely the information that is wanted, without an in-depth knowledge of the SCD system.

Conclusion

The Structural Ceramics Database (SCD) system is being developed as a means of accelerating advances in ceramics-based technology. The first phase of the ongoing SCD program has resulted in a preliminary system for use with personal computers. This phase-one system is focused on the thermal properties of monolithic ceramics. The modular design of the system permits independent modules for materials specification, thermal expansion, thermal conductivity, thermal diffusivity, specific heat, thermal shock resistance, and a bibliography of data references. Accessing the information contained in these modules is accomplished by query and output program elements.

The design of the system has been based on an analysis of three primary considerations relating to the user, the materials properties, and the programming language. Each consideration imposes constraints on the design of the system. The user's interest is the principal determinant of the content of the database and the design of the manner and style with which the system interacts with the user. The programming language determines the technical limitations on how the data are actually managed. The materials properties determine what provisions are necessary to ensure that the critical information is adequately and accurately communicated to the user. The latter provisions impose constraints, for example, on data validation routines for data entry and on query structures for data retrieval.

The success of the SCD system will rest first on its emphasis on user-friendliness and second on its content of critically important data. The convenience, speed, and efficiency of the access to the data will enable developments in research to be transferred to industrial applications far more rapidly than could be expected with the traditional technology transfer processes.

Acknowledgment

The authors wish to thank the generous support of the Gas Research Institute as a founding sponsor in the NIST program to develop the Structural Ceramics Database system. Numerous helpful discussions with C. G. Messina are gratefully acknowledged.

References

[1] Grattidge, W., Westbrook, J., Northrup, C., and Rumble, J., Materials Information for Science and Technology (MIST) Project Overview, Natl. Bur. Stand. (U.S.) Spec. Publ. 726 (1986).

[2] Boerstra, M. L., Engineering Databases (Elsevier, Amsterdam 1985).

[3] Rumble, J., and Sibley, L., Towards a Tribology Information System, Natl. Bur. Stand. (U.S.) Spec. Publ. 737 (1987).

[4] Dabrowski, C. E., and Jefferson, D. K., A Knowledge-Based System for Physical Database Design, Natl. Bur. Stand. (U.S.) Spec. Publ. 500–151 (1988).
[5] Jahanmir, S., Hsu, S. M., and Munro, R. G., ACTIS: Towards a Comprehensive Tribology Database, Proc. of the ASTM Conf. on Computerization and Networking of Material Property Databases (ASTM, Philadelphia 1987).
[6] Kaufman, J. G., Standards Activities of ASTM Committee E-49, First International Symposium on Computerization and Networking of Materials Property Databases (ASTM, Philadelphia 1987).
[7] Onkik, H. M., and Messina, C. G., Creating a Materials Database Builder and Producing Publications for Ceramic Phase Diagrams, Proc. of the ASTM Conf. on Computerization and Networking of Material Property Databases (ASTM, Philadelphia 1987).
[8] National Institute of Standards and Technology, Department of Energy, Department of Defense, National Science Foundation, and Electric Power Research Institute are examples.
[9] American Society for Testing and Materials (ASTM), American Ceramic Society, American Chemical Society, and American Society of Mechanical Engineers are examples.
[10] National Materials Property Data Network, Inc. is an example.
[11] Burlingam, A., Matching a DBMS to User Needs, Mini-Micro Systems, Oct. (1981).
[12] Jones, P. F., Four Principles of Man-Computer Dialogue, Computer Aided Design **10**, 197 (1978).
[13] Stamen, J., and Costello, W., Evaluating Database Languages, Datamation, May (1981).
[14] Dieckman, E. M., Three Relational DBMS, Datamation, Sept. (1981).
[15] Kaufman, J. G., Standards for Computerized Material Property Data Sources and Intelligent Knowledge Systems in Managing Engineering Data: The Competitive Edge, R. E. Fulton, ed. (American Society of Mechanical Engineers, New York 1987).
[16] Hubbard, C. R., Dapkunas, S. J., Munro, R. G., and Hsu, S. M., Advanced Ceramics: A Critical Assessment of Database Needs for the Natural Gas Industry, Natl. Bur. Stand. (U.S.) NBSIR 88-3706 (1988).

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

Applications of the Crystallographic Search and Analysis System CRYSTDAT in Materials Science

Volume 94 Number 1 January–February 1989

T. Siegrist

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

Numerical database systems have recently become available online. Their enhanced search capabilities and fast retrieval of data make them a valuable tool in research. In particular, CRYSTDAT which is a search and analysis system for NBS CRYSTAL DATA has proven to be powerful in the identification of crystalline materials. In conjunction with a single-crystal x-ray diffractometer, a qualitative as well as quantitative phase determination is easily performed. The use of CRYSTDAT will be illustrated in several examples.

Key words: crystallographic database; CRYSTDAT; high T_c superconductors.

Introduction

Numerical database systems compiling inorganic crystallographic data are well established. Among the best known are the JCPDS Powder Diffraction File [1], Structure Reports (Strukturbericht) [2] and Crystal Data Determinative Tables [3]. To retrieve data from the databases in printed form, elaborate and sometimes limited and inconvenient search methods are required. In contrast, online computer systems are highly efficient. CRYSTDAT [4], which is a search and retrieval system for NBS CRYSTAL DATA [5] is accessible online and makes use of efficient search algorithms. It is therefore possible to search many parameter fields efficiently and the retrieved data can be quickly evaluated. In the following, several examples on the possible use of CRYSTDAT will be given. These examples only cover a small part of the different search routines available and use a subset of the options implemented.

Applications

Currently, NBS CRYSTAL DATA contains unit cell, composition, formula, symmetry, and reduced cell data of approximately 115,000 inorganic and organic compounds up to 1985. Using CRYSTDAT, each of these data fields can be searched individually or they can be combined through logical operators. The examples will be restricted to inorganic oxides, illustrating possible uses of CRYSTDAT in the research on high T_c ceramic superconductors. In particular, the first example will show how the database was used in the study of the Ba-Y-Cu-O system. After logging into the account and loading the database, the system replies:

ON-LINE SYSTEM

---------- ----------

DATABASE: CRYSTDAT

START TIME: 20:08:25

DATE: 88-06-16

The number of entries for the CRYSTDAT database=115067
The coverage is for the years 1900–1985

1. Study of the Ba-Y-Cu-O System

>find class i (*find inorganic compounds*)
--Set 1 created with 59951 hits
>set limits 1 (*limits following searches to set 1*)
--Limit Set=1
>find ele O (*find oxides*)
--Set 2 created with 31062 hits
>set limits 2
--Limit Set=2 (*limits following searches to set 2*)

a. Search for Ternary Ba-Cu-Oxides

>find ele Ba.and.Cu.and.3 (*search for ternary barium cuprates*)
--Set 3 created with 3 hits
>show 3 (*present the results of set 3*)

ID[1]: 803667 -- 1
RC: a=5.72 b=5.72 c=6.46 al=116.3 be=116.3 ga=90.0
CD: sys=tetragonal spgr(CD)=I41/amd spno=141 den=6.0(g/cc) z=4
EM: Ba Cu2 O2
FO: Ba Cu2 O2
NM: Barium dicopper(i) oxide
AC: a=5.722 c=10.064 sprg(A)=I41/amd
RF: Z. Naturforsch. B,27,296,1972

ID: 708260 -- 2
RC: a=15.81 b=15.81 c=15.81 al=109.5 be=109.5 ga=109.5
CD: sys=cubic spgr(CD)=I432 spno=211 den=7.6(g/cc) z=0
EM: Ba Cu O2
FO: Ba Cu O2
NM: Barium Copper Oxide
AC: a=18.26 spgr(A)=I432
RF: Rev. Chim. Miner.,13,440,1976

---

[1] Output abbreviations are as follows: ID=compound identification no., RC=reduced cell, CD=Crystal Data, EM=empirical formula, FO=chemical formula, NM=name, AC=author's cell, RF=journal reference.

```
ID : 800744                                                    -- 3
RC : a=15.82 b=15.82 c=15.82 al=109.5 be=109.5 ga=109.5
CD : sys=cubic spgr(CD)=Im3m spno=229 den=5.6(g/cc) z=90
EM : Ba Cu O2
FO : Ba Cu O2
NM : Barium dioxocuprate
AC : a=18.27 spgr(A)=Im3m
RF : Z. Naturforsch. B,32,121,1977
```

b. Search for Ternary Cu-Y-Oxides

```
>find Y.and.Cu.and.3
  --set 4 created with 2 hits
>show 4

ID : 711441                                                    -- 1
RC : a=3.50 b=10.80 c=12.46 al=90.0 be=90.0 ga=90.0
CD : sys=orthorhombic spgr(CD)=Pn21a spno=33 den=5.4(g/cc) z=4
EM : Cu2 O5 Y2
FO : Cu2 Y2 O5
NM : Copper Yttrium Oxide
AC : a=10.799 b=3.4960 c=12.456 spgr(A)=Pna21
RF : 00GRNT,,,1981

ID : 811630                                                    -- 2
RC : a=3.52 b=3.52 c=11.42 al=90.0 be=90.0 ga=120.0
CD : sys=hexagonal spgr(CD)=P63/mmc spno=194 den=5.0(g/cc) z=2
EM : Cu O2 Y
FO : Cu Y O2
NM : Copper(i) yttrium oxide
AC : a=3.5206 c=11.418 spgr(A)=P63/mmc
RF : J. Solid State Chem.,49,232,1983
```

c. Search for Ternary Ba-Y-Oxides

```
>find ele Y.and.Ba.and.3
  --Set 5 created with 4 hits
>show 5

ID : 705238                                                    -- 1
RC : a=3.45 b=10.39 c=12.11 al=90.0 be=90.0 ga=90.0
CD : sys=orthorhombic spgr(CD)=Pnab spno=60 den=7.6(g/cc) z=0
EM : Ba O4 Y2
FO : Ba Y2 O4
NM : Barium Yttrium Oxide
AC : a=10.388 b=12.110 c=3.448 spgr(A)=Pnab
RF : Mater. Res. Bull.,9,1631,1974
```

```
ID : N108141                                                    -- 2
RC : a=3.45 b=10.41 c=12.12 al=90.0 be=90.0 ga=90.0
CD : sys=orthorhombic spgr(CD)=Pnam spno=62 den=5.8(g/cc) z=4
EM: Ba O4 Y2
FO : Ba Y2 O4
NM: Barium yttrium oxide (1˜2˜4)
AC : a=10.415 b=12.120 c=3.455 spgr(A)=Pnam
RF : Z. Naturforsch. B,19,955,1964

ID : 705239                                                     -- 3
RC : a=4.38 b=4.38 c=11.85 al=90.0 be=90.0 ga=90.0
CD : sys=tetragonal den=8.2(g/cc) z=0
EM: Ba2 O5 Y2
FO : Ba2 Y2 O5
NM: Barium Yttrium Oxide
AC : a=4.3771 c=11.852
RF : Mater. Res. Bull.,9,1631,1974

ID : 706037                                                     -- 4
RC : a=6.11 b=6.11 c=25.17 al=90.0 be=90.0 ga=120.0
CD : sys=hexagonal den=2.0(g/cc) z=0
EM: Ba3 O9 Y4
FO : Ba3 Y4 O9
NM: Barium Yttrium Oxide
AC : a=6.1102 c=25.172
RF : Mater. Res. Bull.,9,1631,1974
```

d. And Finally Search for the Quaternary Ba-Y-Cu-Oxides

```
>find ele Ba.and.Y.and Cu.and.4
  --Set 6 created with 1 hits
>show 6

ID : 809291                                                     -- 1
RC : a=5.66 b=7.13 c=12.18 al=90.0 be=90.0 ga=90.0
CD : sys=orthorhombic spgr(CD)=Pbnm spno=62 den=6.2(g/cc) z=4
EM: Ba Cu O5 Y2
FO : Y2 Ba Cu O5
NM: Diyttrium barium copper oxide
AC : a=7.132 b=12.181 c=5.658 spgr(A)=Pbnm
RF : J. Solid State Chem.,43,73,1982
```

Now we can (attempt to) construct a preliminary phase diagram of the Ba-Y-Cu-O system that serves as the starting point of the analysis of this phase space.

e. Identification of a Crystalline Material

In the course of studying the phase system, small black crystals were obtained. From 6 reflections measured on the CAD4 diffractometer, a unit cell of $3.495 \times 6.23 \times 10.795$ Å was inferred. This unit cell serves as the starting point for a database search to identify the phase. In this case, the reduced cell is searched with a preset tolerance to retrieve all phases with similar lattice parameters.

>cells (*calculate the reduced cell and set up the search parameters*)
---Input Cell---
a ? > 3.495
b ? > 6.23 (*if blank, default is the previous value*)
c ? > 10.795
alpha ? > (*the default is 90 °*)
beta ? > (*if blank, default is the previous value*)
gamma ? >
Lattice Type ? (*default P*) >
Tolerance ? (*default 0.10*) >

>Input cell+vol.	3.50	6.23	10.80	90.00	90.00	90.00	235.05
>Lattice type	P						
>Red. Cell+Vol	3.49	6.23	10.79	90.00	90.00	90.00	235.05
>Niggli Matrix	12.215	38.813	116.532				
	0.000	0.000	0.000				
>Metric Symmetry	Orthorhombic						
>Tolerance for cell match	0.100						

Search the data base? (yes/no)>y
--Set 7 created with 0 hits

The database was searched with the reduced cell of the lattice with edges of 3.49, 6.23, and 10.795 Å. No matches were found. However, the unit cell determined may be too small (a subcell) because some reflections could have been missed on the diffractometer. The database is searched again, but this time, possible supercells are included in the search.

>cells
---Input Cell---
a ? > 3.495
b ? > 6.23
c ? > 10.795
alpha ? >
beta ? >
gamma ? >
Lattice Type ? (*default P*) >
Tolerance ? (*default 0.10*) >

>Input cell+vol.	3.50	6.23	10.80	90.00	90.00	90.00	235.05
>Lattice type	P						
>Red. Cell+Vol.	3.49	6.23	10.79	90.00	90.00	90.00	235.05
>Niggli Matrix	12.215	38.813	116.532				
	0.000	0.000	0.000				
>Metric Symmetry	Orthorhombic						
>Tolerance for cell match	0.100						

Search the data base? (yes/no)>n

The database is not searched with the original cell. Instead, a supercell search [6] *will be carried out. CRYSTDAT is instructed to generate possible supercells of 2× the volume of the input cell. After the supercells have been generated, the database is searched.*

Sub-/Super- cells calculated? (YES/NO)
where sub-cell (input cell vol.) / multiplicity
and super-cell (input cell vol.) * multiplicity>y
Sub-cell ? (yes/no)>n
Super-cell ? (yes/no)>y
Multiplicity ? (*2/3/4/5/6/7 ... etc*)>2

Super-cells of 2 times the volume of the input cell

a	b	c	al	be	ga	vol
3.495	6.230	21.590	90.0	90.0	90.0	470
3.495	12.460	12.464	120.0	90.0	90.0	470
6.230	6.990	11.347	107.9	90.0	90.0	470
6.990	7.143	11.347	81.3	72.1	60.7	470
3.495	10.795	12.460	90.0	90.0	90.0	470
6.990	7.143	10.795	90.0	90.0	119.3	470
6.230	6.990	10.795	90.0	90.0	90.0	470

Search the data base? (yes/no)>y

NOW SEARCHING

The database is now searched including all the above given supercells of multiplicity 2.

--Set 8 created with 31 hits

In this set of retrieved entries, we find the following one:

ID : 711441 -- 1
RC : a=3.50 b=10.80 c=12.46 al=90.0 be=90.0 ga=90.0
CD : sys=orthorhombic spgr(CD)=Pn21a spno=33 den=5.4(g/cc) z=4
EM: Cu2 O5 Y2
FO : Cu2 Y2 O5
NM: Copper Yttrium Oxide
AC : a=10.799 b=3.4960 c=12.456 spgr(A)=Pna21
RF : 00GRNT,,,1981

Supercell 5 from the above list of supercells matched a cell in the database of a compound with the expected elemental composition. In this way, both the correct lattice and an accurate composition of the small black crystals were determined in roughly 30 minutes.

2. Study of the Bi-Sr-Ca-Cu-O System

In the case of the Bi-Sr-Ca-Cu-O superconductors, the analogous searches are easily done.

a. Search for Ternary Bi-Cu Oxides

>find ele Bi.and.Cu.and.3
--Set 9 created with 4 hits
>show 9

ID : B034138 -- 1
RC : a=5.80 b=8.48 c=8.48 al=90.0 be=90.0 ga=90.0
CD : sys=tetragonal spgr(CD)=P4/ncc spno=130 den=8.0(g/cc) z=2
EM: Bi4 Cu O7
FO : Bi4 Cu O7
NM: Bismuth copper oxide (4~1~7)
AC : a=8.481 c=5.803 spgr(A)=P4/ncc
RF : An. Acad. Brasil. Cienc.,38,35,1966

ID : 804574 -- 2
RC : a=5.81 b=6.67 c=6.67 al=79.0 be=64.1 ga=64.1
CD : sys=tetragonal spgr(CD)=I4 spno=79 den=8.6(g/cc) z=4
EM: Bi2 Cu O4
FO : Cu Bi2 O4
NM: Copper dibismuthate(iii)
AC : a=8.484 c=5.813 spgr(A)=I4
RF : Z. Anorg. Allg. Chem.,426,1,1976

ID : 704311 -- 3
RC : a=5.81 b=8.51 c=8.51 al=90.0 be=90.0 ga=90.0
CD : sys=tetragonal spgr(CD)=P4/ncc spno=130 den=8.6(g/cc) z=4
EM: Bi2 Cu O4
FO : Cu Bi2 O4
NM: Copper Bismuth Oxide
AC : a=8.510 c=5.814 spgr(A)=P4/ncc
RF : C. R. Hebd. Seances Acad. Sci. Ser. C,276,1105,1973

ID : 810281 -- 4
RC : a=5.81 b=8.51 c=8.51 al=90.0 be=90.0 ga=90.0
CD : sys=tetragonal spgr(CD)=P4/ncc spno=130 den=8.6(g/cc) z=4
EM: Bi2 Cu O4
FO : Cu Bi2 O4
NM: Bismuth copper oxide
AC : a=8.510 c=5.814 spgr(A)=P4/ncc
RF : Bull. Soc. Fr. Mineral. Cristallogr.,99,193,1976

b. Search for Ternary Bi-Sr-Oxides

>find ele Bi.and.Sr.and.3
--Set 10 created with 4 hits
>show 10

ID : B028466 -- 1
RC : a=3.95 b=3.95 c=9.64 al=78.2 be=78.2 ga=60.0
CD : sys=rhombohedral den=6.0(g/cc) z=0
EM: Bi38 O59 Sr2
FO : Bi38 Sr2 O59
NM: Bismuth strontium oxide (38~2~59)
AC : a=3.952 c=28.09 spgr(A)=R
RF : J. Res. Nat. Bur. Stand. Sect. A,68,197,1964

ID : 805401 -- 2
RC : a=3.97 b=3.97 c=9.75 al=78.2 be=78.2 ga=60.0
CD : sys=rhombohedral spgr(CD)=R-3m spno=166 den=7.8(g/cc) z=9
EM : Bi0.76 O1.38 Sr0.23
FO : Bi0.765 Sr0.235 O1.383
NM : Bismuth strontium oxide (.8/.2/1.1)
AC : a=9.75 al=23.49 spgr(A)=R-3m
RF : J. Solid State Chem.,35,192,1980

ID : 812013 -- 3
RC : a=3.97 b=3.97 c=9.74 al=78.2 be=78.2 ga=60.0
CD : sys=rhombohedral spgr(CD)=R-3m spno=166 den=7.8(g/cc) z=4
EM : Bi1.72 O3 Sr0.53
FO : Bi1.72 Sr0.53 O3
NM : Bismuth(iii) strontium(13/4) oxide
AC : a=3.971 c=28.41 spgr(A)=R-3m
RF : Solid State Ionics,3,457,1981

ID : 709884 -- 4
RC : a=4.26 b=9.60 c=9.60 al=87.2 be=77.2 ga=77.2
CD : sys=tetragonal spgr(CD)=I4/m spno=87 den=7.2(g/cc) z=0
EM : Bi1.10 O2.55 Sr0.90
FO : Sr0.9 Bi1.1 O2.55
NM : Strontium Bismuth Oxide
AC : a=13.239 c=4.257 spgr(A)=I4/m
RF : Rev. Chim. Miner.,15,153,1978

c. Search for Bi-Cu-Sr or Bi-Cu-Ca Compounds with 4 or 5 Elements

>find ele Bi.and.Cu.and.Sr.and.5
--Set 11 created with 0 hits
>find ele Bi.and.Cu.and.Sr.and.4
--Set 12 created with 0 hits
>find ele Bi.and.Ca.and.Cu.and.4
--Set 13 created with 0 hits
>find ele Bi.and.Cu.and.Ca.and.5
--Set 14 created with 0 hits

No entries containing 4 and 5 elements were found.

d. Find materials with lattices related to that of a superconducting crystal

From the x-ray study of a superconducting crystal, we obtained a cell of approximately 5.44 × 5.43 × 30.8 Å. This cell data can be used to retrieve compounds that may be structurally related to the superconductors. For this purpose, tetragonal or orthorhombic cells are retrieved and searched for their c-axis values given by the authors.

Carry out several searches to generate SET 17 containing inorganic materials with Bi and O; restrict next search to this set.

>Find sys O.or.T (*orthorhombic or tetragonal systems only*)
--Set 18 created with 244 hits
>Set limits 18 (*limits now set to inorganic bismuth oxides belonging to the orthorhombic or tetragonal systems.*)

--Limit Set=18

>find cc 28.to.34 (*find compounds with the author's c-axis between 28 and 34* Å)

--Set 19 created with 8 hits

In this way, compounds with unit cell c-axes between 28 to 34 Å *are retrieved more efficiently than by using the "cells" command.*

>show 19

ID : 711644 -- 1
RC : a=3.83 b=3.83 c=33.62 al=90.0 be=90.0 ga=90.0
CD : sys=tetragonal den=7.4(g/cc) z=0
EM: Bi3 F O11 Pb Ti3
FO : Pb Bi3 Ti3 O11 F
NM: Lead Bismuth Titanium Oxide Fluoride
AC : a=3.833 c=33.62 spgr(A)=P4/mm
RF : J. Solid State Chem.,36,349,1981

ID : B019359 -- 2
RC : a=3.84 b=3.84 c=16.64 al=96.6 be=96.6 ga=90.0
CD : sys=tetragonal spgr(CD)=I4/mmm spno=139 den=8.0(g/cc) z=2
EM: Bi4 O12 Ti3
FO : Bi4 Ti3 O12
NM: Bismuth titanium oxide (4^3^12)
AC : a=3.841 c=32.83 spgr(A)=I4/mmm
RF : Ark. Kemi,1,499,1949

ID : B024961 -- 3
RC : a=3.92 b=3.92 c=14.71 al=97.7 be=97.7 ga=90.0
CD : sys=tetragonal den=7.2(g/cc) z=2
EM: Bi3 Br3 Ca O4
FO : Bi3 Ca Br3 O4
NM: Bismuth calcium bromide oxide (3^1^3^4)
AC : a=3.92 c=28.90 spgr(A)=I
RF : Z. Anorg. Allg. Chem.,248,121,1941

ID : B024960 -- 4
RC : a=3.98 b=3.98 c=14.66 al=97.8 be=97.8 ga=90.0
CD : sys=tetragonal spgr(CD)=I4/mmm spno=139 den=7.4(g/cc) z=2
EM: Bi3 Br3 O4 Sr
FO : Bi3 Sr Br3 O4
NM: Bismuth strontium bromide oxide (3^1^3^4)
AC : a=3.976 c=28.78 spgr(A)=I4/mmm
RF : Z. Anorg. Allg. Chem.,246,115,1941

ID : B024962 -- 5
RC : a=4.04 b=4.04 c=16.23 al=97.2 be=97.2 ga=90.0
CD : sys=tetragonal spgr(CD)=I4/mmm spno=139 den=7.4(g/cc) z=2
EM: Bi3 I3 O4 Sr
FO : Bi3 Sr I3 O4
NM: Bismuth strontium iodide oxide (3^1^3^4)
AC : a=4.043 c=31.95 spgr(A)=I4/mmm
RF : Z. Anorg. Allg. Chem.,250,173,1942

ID : 709204 -- 6
RC : a=5.40 b=5.44 c=29.05 al=90.0 be=90.0 ga=90.0
CD : sys=orthorhombic den=3.0(g/cc) z=0
EM: Bi7 Nb O21 Ti4
FO : Bi7 Ti4 Nb O21
NM: Bismuth Titanium Niobium Oxide
AC : a=5.44 b=5.40 c=29.05 spgr(A)=P63
RF : J. Less-Common Metals,48,319,1976

ID : 805515 -- 7
RC : a=5.41 b=5.45 c=16.64 al=99.4 be=90.0 ga=90.0
CD : sys=orthorhombic spgr(CD)=C2ca spno=41 den=8.0(g/cc) z=4
EM: Bi4 O12 Ti3
FO : Bi4 Ti3 O12
NM: Tetrabismuth trititanium oxide
AC : a=5.448 b=5.411 c=32.83 spgr(A)=B2cb
RF : Ferroelectrics,3,17,1971

ID : N110983 -- 8
RC : a=5.41 b=5.45 c=32.82 al=90.0 be=90.0 ga=90.0
CD : sys=orthorhombic den=8.0(g/cc) z=4
EM: Bi4 O12 Ti3
FO : Bi4 Ti3 O12
NM: Bismuth titanium oxide (4~3~12)
AC : a=5.411 b=5.449 c=32.82
RF : J. Electrochem. Soc.,116,832,1969

As it turns out, the last entry served as a starting point for the successful refinement of the Bi-Sr-Ca-Cu-O superconductor phase.

The set of commands a user has to remember is small and the commands themselves are self explanatory making CRYSTDAT an efficient tool in materials research.

References

[1] The Powder Diffraction File, JCPDS—International Centre for Diffraction Data, Swarthmore, PA 19081.
[2] Structure Reports, Published for the International Union of Crystallography, Vols. 8–46A (1940–1981), Reidel, Dordrecht.
[3] Crystal Data Determinative Tables, third edition, Vol. 1 (1972), Vol. 2 (1973), Vols. 3–4 (1978), Vols. 5–6 (1983). U.S. Department of Commerce, National Bureau of Standards, and the JCPDS—International Centre for Diffraction Data, Swarthmore, PA.
[4] CRYSTDAT (1987). An Online Search and Analysis System for NBS CRYSTAL DATA. CRYSTDAT was developed jointly by CISTI's (Canada Institute for Scientific and Technical Information) CAN/SND Scientific Numeric Database Service and the NIST Crystal Data Center.
[5] NBS CRYSTAL DATA (1987). A Magnetic Tape of Crystallographic Data Compiled by the NIST Crystal Data Center, National Institute of Standards and Technology, Gaithersburg, MD.
[6] Mighell, A. D., and Himes, V. L., Compound Identification and Characterization Using Lattice-Formula Matching Techniques, Acta Cryst. **A42**, 101 (1986).

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

New Directions in Bioinformatics

Volume 94 Number 1 January–February 1989

Daniel R. Masys

Lister Hill National Center
for Biomedical Communications
National Library of Medicine
Bethesda, MD 20894

Two decades have passed since the first large scale, public access computer-based information systems were developed to store and disseminate the knowledge of medicine and biology. These first systems were bibliographic, and though the searching of computer files of citations remains the most common use of biological databases, there are dramatic forces at work in basic biology which are driving a transition from the printed page to the factual database. Unlike bibliographic systems, which contain only a pointer to information located elsewhere, factual databases contain the information sought. Development of automated methods to sequence DNA, RNA, proteins, and other macromolecules have yielded oceans of cryptic symbols, for which there is an absolute dependence upon computerized factual databases to acquire, store, retrieve, and analyze data. The Human Genome Project has focussed attention on the information science aspects of nucleic acid data, yet for the practicing scientist nucleic acids and other sequence data are just one piece of an increasingly complex biological puzzle whose solution will be expressed in terms of structure and function. Access to and integration of information across multiple related biological databases is a major challenge facing information system builders, a challenge which holds the promise of creating knowledge synergy from what are today disconnected, stand-alone information sources.

Key words: biotechnology; computer communication networks; computer systems; database management systems; medical informatics; molecular biology; National Library of Medicine (U.S.).

Background

The methods for acquiring, storing, manipulating, and disseminating scientific knowledge are changing rapidly. Though the printed page remains the principal medium for the publication of research results, in data intensive applications such as crystallography, the advantages of computers as generalized symbol processors are leading to a paradigm shift from the printed page to the factual database. Among the public institutions most profoundly affected by this change is the National Library of Medicine (NLM). NLM celebrated its 150th anniversary in 1986, honoring a long tradition as the world's largest repository and distribution center for the published knowledge of medicine and biology. In 1971, the Library initiated the MEDLARS online bibliographic files including MEDLINE and now over 20 other databases. MEDLINE remains the most widely used bibliographic file in biomedicine, answering some 4 million online inquiries this past year.

The Lister Hill National Center for Biomedical Communications is the Research and Development Division of the Library, serving as an intramural laboratory analogous to the intramural labs of the other institutes of the NIH. The Lister Hill Center has a staff of about 80, many of whose professional

backgrounds are a hybrid of biological or medical science and computer and information science. The Lister Hill Center conducts research and development in three major areas: Computer and Information Science as applied to biomedical research and health care delivery; Electronic Image technologies such as image capture, compression, storage on optical disc, retrieval and transmission; and Health Professions Education using new information technologies such as microcomputers and videodiscs.

Biotechnology Information

A new era in the mission of the National Library of Medicine was inaugurated with the appointment of Dr. Donald Lindberg as its Director in 1984. One of his first strategic moves was, along with the Library's Board of Regents, to commission a long range plan for the Library. Calling together some 120 experts in the fields of library science, computer and information science, and health professions education, he charged the Long Range Planning panels with picturing the future 5, 10, and 20 years hence, and made recommendations to move the Library into that information technology-rich future. Of the more than 80 recommendations made to the Library, that given the highest institutional priority was the development of better information systems in what has come to be known as Biotechnology [1].

Biotechnology information refers to the computerized factual databases of molecular and structural biology which are growing rapidly in size and number. In the setting of the growing Human Genome Project, the current central focus of these research databases is DNA in partucular, and sequence databanks in general. The development of automated methods to analyze DNA, RNA, and protein sequences has, for the first time in the history of biology, brought to a large fraction of scientists an absolute dependence upon computers for the representation, storage, retrieval, and analysis of this data which far exceeds the limits of human cognition to remember, to scan, and to detect patterns of significance.

There are a number of conceptually related but technically dissimilar databases which are national resources for molecular biology research, and most are supported by the NIH in one form or another (fig. 1). In the aggregate, they might be viewed as an electronic Tower of Babel, a maze of contrasting record structures and searching dialects which frustrates all but the most ardent researcher who wishes to access more than one of them. During the past 2 years, NLM has listened to the scientists, looked at its own institutional strengths, and has begun the work of devising better information management systems and methods for the knowledge of molecular and structural biology.

Current NLM Biotechnology Activities

The NLM's current activities fall into three major areas: the building and maintenance of databases relevant to molecular biology; new methods of retrieval and analysis of the information in those databases; and education, both for the researchers who generate the data and need to familiarize themselves with the computer-based tools being developed to facilitate management and analysis of the information, and for computer scientists who need to understand the data analysis problems of biologists.

In the area of database building and maintenance, several new computer index terms have been created which denote that an article contains nucleic acid or protein sequence information. The indexing coverage of the NLM will provide surveillance of nearly 3000 journals for the appearance of new literature in this area. As indexers apply the new tags to the literature, MEDLINE subsets will be made available in computer readable form to molecular biology database builders, to use either as a current awareness service for papers appearing in journals that they do not normally scan, or even as the kernel for new entries in their databases. Virtually every one of the factual databases of molecular biology has a bibliographic citation as a component of its unit record.

The MEDLINE searching vocabulary and computer record structures are being modified to enable automated linkage between the published literature and the DNA research databases. Once these links are in place, it will be possible for an investigator who is browsing through the DNA database to tell the computer to find literature related to one or a group of GenBank records. Going the other way, the computer will allow those who are doing literature searches in molecular biology to automatically connect to the databases where the actual data cited in the publication is contained. Although the DNA databases are the first prototype linkage, we hope to extend this interlinkage to all of the major biology databanks, including biophysical and structural databanks such as Brookhaven and Cambridge.

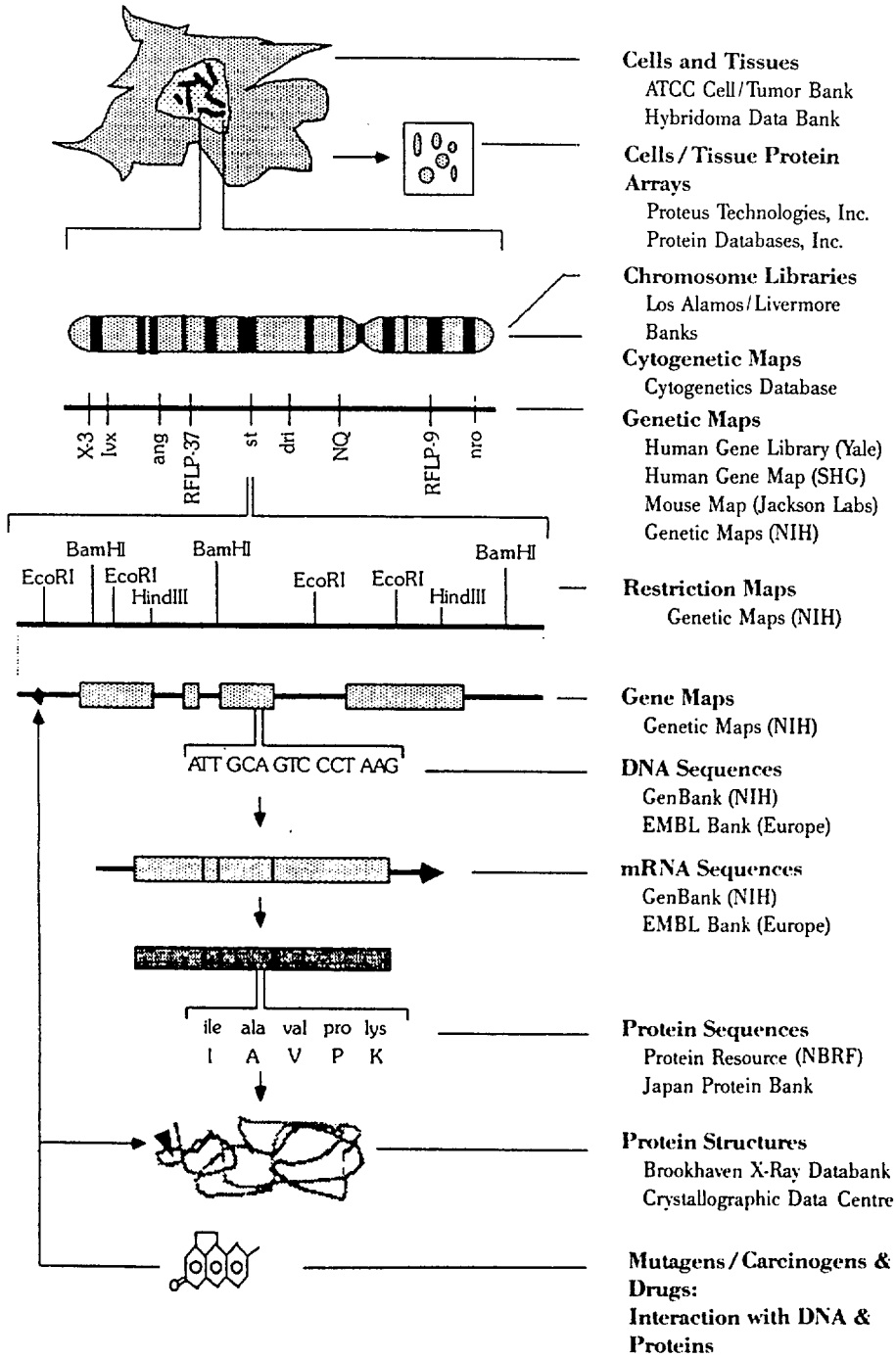


Figure 1. Biology knowledge bases.

Other new database building efforts include a database of databases, that is, a Directory of Biotechnology Information Resources. This collection will become an online file, so that if an investigator wants to know where a particular biotechnology database is located, and how to connect to it, he can browse our online catalog of research databases.

One important social and political aspect of biotechnology is the issue of environmental release of genetically altered organisms. At the request of the U.S. Biotechnology Science Coordinating Committee, the NLM is working with representatives from domestic U.S. agencies such as the EPA, Dept. of Agriculture, and NSF, and foreign delegates from the Commission of European Commu-

nities, and various European scientific databases to develop an environmental release database which will serve as both a scientific and regulatory information resource.

Since 1982, the Lister Hill Center has been supporting the development of Victor McKusick's "Mendelian Inheritance in Man," the acknowledged world standard reference describing over 4000 human genetic diseases. We have converted the entire textbook into a full text, online database, searchable by natural language query. Current efforts underway are enhancing the text with a library of images, including x rays and clinical photos of patients with the genetic diseases described in the text.

The building of new and more capable electronic systems for collecting and storing biotechnology information is not enough in itself to ensure progress. We need better ways of retrieving and analyzing that information, in a manner that is easy for both expert and novice users. The concept of simultaneous retrieval of related information from multiple databases is central to development work now underway in the Lister Hill Center's Information Technology Branch. Our Information Retrieval Experiment, or IRX software, is being adapted to the special needs of biotechnology databases.

The system model is based on a windowing scientific workstation which allows natural language questions to be asked, with retrieval of related information from a number of conceptually related databases. The prototype system has simultaneous access to 12 different databases, including the Brookhaven crystal structures. New user interfaces based on high resolution windowing workstations, are a part of this work. A pseudo-natural language allows the user to state a question in his own language, such as "Has Factor XII deficiency been mapped?" The keywords of the query are parsed and truncated to form word stems, then all occurrences of those word stems are statistically ranked to generate an ordered retrieval set.

The new and rapidly changing knowledge of molecular and structural biology confers an ongoing need for education, and the Library has instituted a series of biotechnology lectures. These introductory lectures on the topics of the new genetics and the information science aspects of the field have been put on videotape and are available as part of NLM's lending collection. Educational efforts in computer and information science as applied to molecular biology need also to be directed to the workers in the field. NLM has hosted more than 60 NIH molecular biologists for workshops on computerized "Molecular Sequence Analysis".

The Future in Computational Biology

It is clear that molecular and structural biology is an area of science which, perhaps more than any other component of biology or medicine, has acquired an absolute dependence upon computers to carry forward the advancement of science. At the NLM, the first major objective of the future will be building, maintaining, and providing access to research information resources, in much the same way that the Library organizes, indexes, and provides access to the scientific literature. The second objective will be support for scientific discovery.

Information Resources programs will include ongoing support for key molecular databases. NLM will continue to be one of many distribution channels for this kind of research information, using the Library's extensive international MEDLARS network, and importantly, NLM will focus at the interfaces between these biology knowledge sources, and help develop and promote standards which will permit them to exchange data and provide access to conceptually related information located in multiple databases.

It is clear that the information resources will continue to evolve optimally only if they follow a vanguard of informatics research (i.e., scientists, who use the databanks as the substrate for experimentation and hypothesis testing and who can paint the scientific concepts which data structures must accomodate, must actively collaborate with those responsible for the development and application of the information resources). To this end, a focus of the Biotechnology Information Center Program is a core intramural research staff of hybrid computer scientists and biological scientists, complemented, as it is in the other intramural laboratories of the NIH, by a visiting scientist program. Grant support for the computational aspects of molecular and structural biology would be a substantial component of the program, as would the convening of tutorials, workshops, and scientific meetings highlighting the union of computer science and biological science.

These are exhilarating times in biology and also in information systems development; the National Library of Medicine will be pleased to work with each of the laboratories and institutions represented

at this conference to improve the information tools of basic biology and their inevitable effect on the future practice of molecular medicine.

Reference

[1] National Library of Medicine. Long range plan. Report of the Board of Regents. Bethesda, MD: U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health; 7 vols. (1986–87).

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

The Use of Structural Templates in Protein Backbone Modeling

Volume 94 Number 1 January–February 1989

Lorne S. Reid

Allelix Biopharmaceuticals
6850 Goreway Drive
Mississauga, Ontario
Canada, L4V 1P1

The procedures used to model a protein structure are well established when the novel protein has high sequence similarity to a protein of known structure. Many proteins of interest have low (i.e. $<50\%$) sequence similarity to any known structure. In these cases new approaches to prediction of structure are required.

The use of sequence profiles which relate sequence to known structure has been proposed as one method to assign local regions of structure. As a first stage, templates or "icons" of the many relevant substructural motifs found in proteins must be defined. The sequences which gave rise to these structures are then aligned and a weighted profile obtained.

Average structures of the 8 and 12 residue helix-turn and turn-helix motifs have been prepared. These coordinate templates were then used to scan through the Brookhaven protein structural database for similar, superimposable fragments. A composite template of 100 similar fragments for each element was found to be internally consistent to a rmsd = 0.92 Å for HT8, 1.54 Å for HT12, 0.41 Å for TH8 and 1.40 Å for TH12. All of the sequences, from these structures, were then used to create an overall sequence profile.

The four sequence profiles were scanned against the amino acid sequences of the proteins in the Brookhaven database: tertiary structure was correctly identified only about 10% of the time. This value is too low for predictive purposes. However, it could be increased by checking for multiple occurrences of the template in one protein.

Key words: α helix; β turn; compact domains; modeling; protein structure; sequence profiles; structure prediction; templates.

1. Introduction

The process of protein modeling relies upon the database of structures determined principally by x-ray crystallography or, more recently, 2-D NMR techniques. As a first step in modeling, the degree of sequence similarity of a novel protein is compared to all proteins of known structure. Given high sequence similarity ($>50\%$) the techniques of homology modeling will certainly be used [1–7]. The effectiveness of this process has been demonstrated in the construction of models of insulin-like growth factor [8], t-PA [9], and immunoglobulin variable domain [10] to name a few. However, many proteins of interest have a lower degree of homology or obvious insertions or deletions in their sequence. Any methods which can be used to predict the structure of these proteins are of great interest to experimentalists and theoreticians alike.

The secondary structure of a protein can be predicted with methods such as Chou-Fasman but only to some 65% accuracy [11,12]. To improve upon this, the use of sequence specific profiles has been proposed [1,13,14]. The sequence specific requirements of β turns [15], N-cap, C-cap α helices [16] and proline-kinked α helices [17] have been

previously defined. Also, the sequence requirements of large domains are known for the globin fold [18,19], and the immunoglobulin fold [20].

A major assumption in this procedure is that certain linear amino acid sequences give rise to specific structural elements [21–23]. Many different approaches have been taken to identify zones in proteins which are very closely packed [24–32]. Most methods are computationally intensive; one simple method is to count the number of residues which lie within a sphere of a given radius around any atom. To prepare a profile, the relevant fragments are extracted from all proteins of known structure and aligned in space. The amino acid types are then checked at each residue position and a weighted sequence profile determined. Any novel amino acid sequence can then be checked against a bank of such known profiles and the most likely tertiary fragments identified. This procedure differs from the standard predictive methods of secondary structure in that it attempts to assign specific three-dimensional structure on the basis of sequence and not just regions of secondary structure.

In this work, two examples of both turn-helix and helix-turn structures were chosen for study. These structures were previously identified by Zefus as highly compact structures which were repeated throughout many protein structures [31]. The purpose of this work is to outline some of the steps involved in the identification of relevant templates and their application to structure prediction.

2. Methodology

All programs were written in Fortran 77 and run on a VAX 11/750 under the VMS rev 4.7 operating system.

2.1 Preparation of Stage I Templates

The number and identity of residues which surround each residue in the protein lysozyme (Brookhaven code 1LZ1) were determined. The radius of the sphere checked around each atom was over the range of 3.0 to 8.0 Å.

2.2 Identification of Average Structural Template Coordinates

For the purposes of this work four structural units of a known compact nature were used. These were the 8 residue helix-turn (HT8), 12 residue helix-turn (HT12), 8 residue turn-helix (TH8), and 12 residue turn-helix (TH12) domains as assigned by Zefus [31].

2.2.1 Preparation of Stage I Templates The backbone coordinates of each member associated with a structural template were superimposed using a conjugate gradient rotation/translation function. The root mean square deviation (rmsd) of each member to every other member was calculated for both the main-chain and side-chain atomic positions.

If a particular member appeared to be significantly different from all the other members it was discarded from further consideration. The mean X, Y, Z coordinates of the main-chain atoms were calculated from the fragments under consideration. This coordinate set was identified as a stage I template.

2.2.2 Preparation of Stage II Templates Only proteins in the Brookhaven database (release October 1987) with a resolution of better than 2.5 Å were used in this work [33]: 82 non-homologous proteins, 177 proteins in total were used in this subset of the database. The 100 fragments with the lowest rmsd to the stage I template were rank ordered and the average coordinate set calculated. Finally, the average of the standard deviation of the errors in the X, Y, and Z coordinates was determined. This new coordinate set was identified as stage II template.

2.3 Amino Acid Sequence Profiles

The amino acid sequences used to prepare the stage II template were assembled with the programs of the University of Wisconsin Genetics Computer Group (Ver 5.2) [34]. A sequence profile was prepared with the program PROFILE [13]. The Protein Identification Resource/NBRF (PIR) (Rel 15.0) database [35] of amino acid sequences was scanned with the program PROFILESEARCH and alignments calculated with PROFILESEGMENTS. A subset of the PIR database, which corresponded to the proteins used in the Brookhaven database, was also checked for alignments to the calculated profiles.

3. Results

For the purposes of modeling or structure prediction it is necessary to clearly define substructural elements. A number of canonical structures such as α helices, β sheets or larger super-

secondary elements such as Greek keys, or α-β-α units are well known. However, irregular or compound elements can have a very high packing density. Inter-residue contact plots are a convenient method for identification of both the contiguous and discontinuous zones of high density (data not shown).

The number of contacts which a particular residue makes with its neighbors increases in a linear way with the size of the probe distance [36]. As shown in figure 1 for lysozyme (1LZ1) beyond a shell size of 4.0 Å the shape of the compact domain did not change; there was an increase only in the number of residues involved. Two of the structural templates under investigation exist in the lysozyme structure and occur in regions of high packing density. Neither of the motifs in lysozyme were used to generate the stage I templates.

The fragments used for the preparation of stage I templates are given in table 1. A number of elements originally identified by Zefus as compact turn-helix 8 motifs were rejected for use in the preparation of the stage I TH8 template. Rejection was based upon an average rmsd, of the fragment to all other members of the test set (main-chain atoms only), of 1.5 Å greater than the average rmsd for all residues in the $N \times N$ test set.

Table 1. Residues used in the generation of stage I templates

Helix-turn 8		Helix-turn 12		Turn-helix 8		Turn-helix 12	
Range	File[a]	Range	File	Range	File	Range	File
6- 13	2ACT	35- 46	2ACT	98-105	2ACT	19- 30	2ACT
75- 82	2ACT	122-133	2ACT	13- 20	5CPA	89-100	5CPA
116-123	5CPA	227-238	5CPA	95-102	4DFR	97-108	3CPV
242-249	5CPA	255-266	5CPA	38- 45	3FXN	90-101	3CYT
28- 36	3CPV	99-110	3FXN	92- 99	3FXN	105-116	6LYZ
9- 16	3CYT	142-153	3MBN	3- 10	6LYZ	45- 56	4PTI
31- 38	6LYZ	35- 46	8PAP	78- 85	6LYZ	2- 13	5RSA
92- 99	3MBN	119-130	8PAP	2- 9	3MBN	297-308	3TLN
6- 13	8PAP	98-109	2SNS	99-106	3MBN		
73- 80	8PAP			123-130	3MBN		
14- 21	1SBT			1- 8	4PTI		
147-154	3TLN						
240-247	3TLN						
268-275	3TLN						
Average rmsd of superimposed main-chain atomic coordinates (Å)							
1.79±0.54[b]		2.67±1.15		1.08±0.40		2.80±0.86	
Average rmsd of superimposed side-chain atomic coordinates (Å)[c]							
2.13±0.83		3.85±1.61		1.72±0.65		3.92±1.13	
Average number of side-chain atoms superimposed over the entire template							
12.0±3.8		16.8±3.6		12.9±4.0		17.3±6.7	

[a] Brookhaven code.
[b] Error expressed as standard deviation.
[c] Side-chain coordinates were checked between superimposed structures if their atomic name was the same.

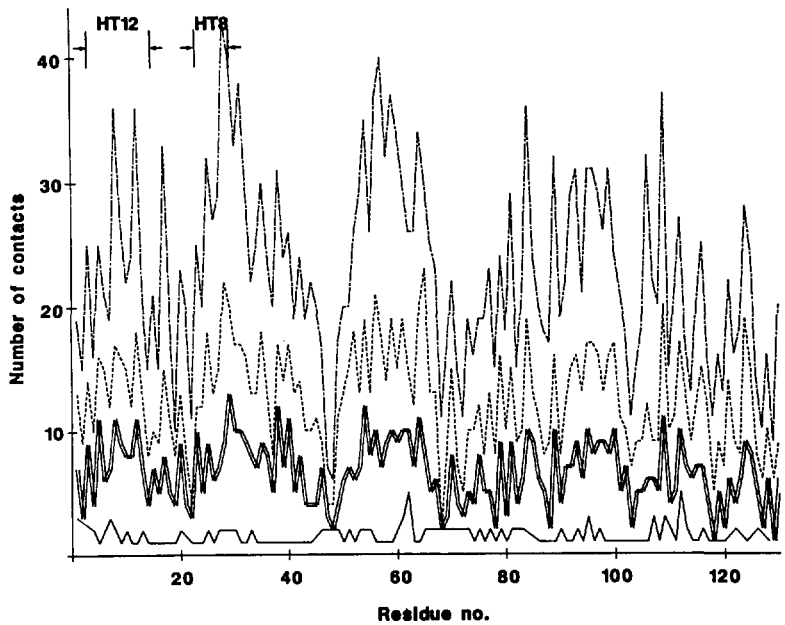


Figure 1. Nearest neighbor contacts in lysozyme (1LZ1) as a function of interatomic distance: 3.0 Å (——), 4.0 Å (═══), 6.0 Å (– – – –), 8.0 Å (—·—·). The TH12 and TH8 motifs exist in the protein at the identified regions of high packing density.

Superimposition of the coordinate sets was based solely upon the backbone atoms. Those side-chain atoms which had equivalent atom names at superimposed residues were checked for structural homology. For example, if the backbones of alanine and cystine were superimposed the rmsd was determined for the $C\beta$ atom position. On average, 1.5 side-chain atomic positions could be superimposed at each residue over all the paired coordinate sets.

The turn-helix 8 stage I template had the greatest degree of structural homology for both main-chain and the superimposable side-chain atoms. In each stage I template the greatest diversity occurred in the turn region: the helix was well defined. This may relate to actual differences in the structure and partly to the difficulty of building the original protein structure into x-ray density associated with irregular elements such as these turns. Alternatively, this may indicate that average rmsd error is a relatively insensitive indicator of similarity between protein fragments.

The Brookhaven protein database was scanned for the best 100 fragments which could be superimposed onto the stage I template. Due to the existence of multiple forms and multiple chains in a protein the database has significant redundancy. However, these redundant fragments had minor variations in three dimensional structure. Keeping and averaging these redundant forms reduced the structural error associated with the motif as found in any one particular crystal structure. Table 2 indicates the average rmsd values of the top 50 and top 100 fragments which were found in this manner for each template type.

Table 2. rmsd of fragments extracted from the Brookhaven database to stage I coordinates

Template	Top 50 fragments rmsd (Å)	±[a] (Å)	Top 100 fragments rmsd (Å)	± (Å)
Helix-turn 8	0.85	0.04	0.92	0.08
Helix-turn 12	1.45	0.09	1.54	0.12
Turn-helix 8	0.38	0.02	0.41	0.03
Turn-helix 12	1.36	0.03	1.40	0.05

[a] Error expressed as standard deviation.

The average structure of the HT8 stage II template is shown in figure 2, HT12 in figure 3, TH8 in figure 4 and TH12 in figure 5. The sphere centered at each atom represents 50% of the standard deviation error in atomic position at that atom between all members used to generate the stage II template. The templates were relatively structurally homologous. The helix atoms in both 8 residue tem-

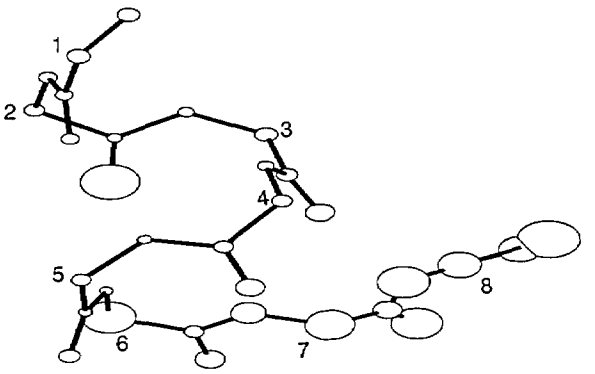


Figure 2. Helix-turn 8 residue stage II template. Sphere size represents 50% of the rmsd error at each atomic position. The Cα atom of each residue is numbered. Picture generated by the PLUTO program.

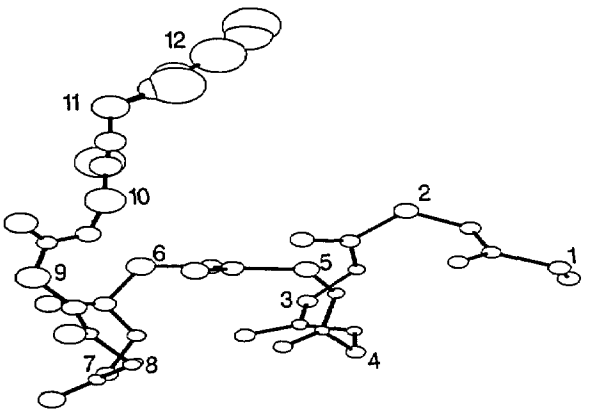


Figure 3. Helix-turn 12 residue stage II template. Sphere size represents 50% of the rmsd error at each atomic position.

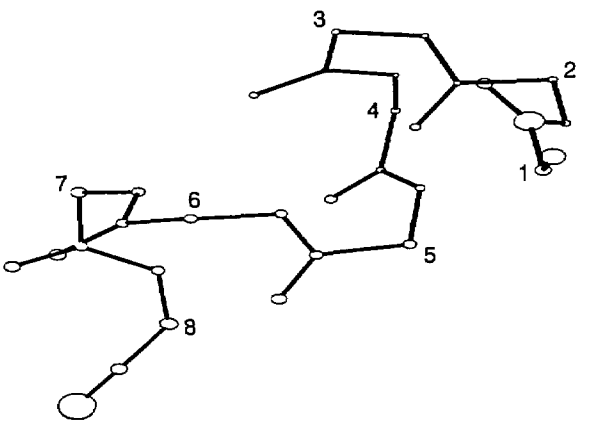


Figure 4. Turn-helix 8 residue stage II template. Sphere size represents 50% of the rmsd error at each atomic position.

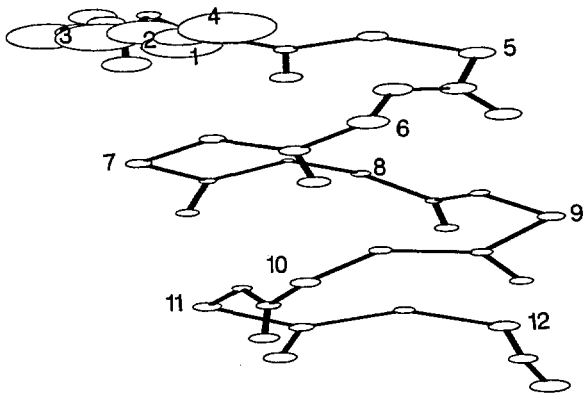


Figure 5. Turn-helix 12 residue stage II template. Sphere size represents 50% of the rmsd error at each atomic position.

plates had an error (0.30±0.1 Å) close to the experimental error of the protein coordinate sets whereas the atoms associated with the turn were less well defined (0.4±0.2 Å). The longer 12 residue templates were less accurate with an average error of 0.7±0.3 Å in the turn regions, double that of the helix region (0.3±0.1 Å). The associated *X*, *Y*, *Z* coordinates are given in Appendix 1: phi, psi backbone angles of each template are given in table 3. Residues in the turn did not correspond to any of the standard β turn types.

Table 3. Backbone phi, psi angles of the stage II templates

Residue no.	Helix-turn 8		Helix-turn 12		Turn-helix 8		Turn-helix 12	
	Phi	Psi	Phi	Psi	Phi	Psi	Phi	Psi
1		−41.0		−41.9		144.0		40.9
2	−65.5	−41.7	−61.9	−40.1	−72.7	151.2	−111.6	16.5
3	−65.0	−37.0	−62.3	−40.6	−55.4	−37.6	−87.5	−146.1
4	−71.6	−43.5	−63.7	−43.2	−62.1	−39.8	−58.6	−44.4
5	−74.7	−35.4	−62.8	−40.1	−69.1	−37.0	−64.6	−43.0
6	−99.3	−15.3	−63.6	−38.6	−66.5	−39.6	−66.7	−39.5
7	93.3	53.4	−65.2	−30.8	−66.5	−35.1	−60.1	−43.7
8			−88.7	−15.1			−62.5	−41.8
9			107.2	23.5			−64.6	−44.0
10			−106.4	164.6			−64.7	−40.7
11			−84.5	149.8			−62.0	−41.2
12								

The sequences of the top 100 residues used to generate the stage II template were compiled and subjected to PROFILE analysis. The profiles are given in Appendix 2, consensus sequences are shown in table 4. Standard weighting, a gap penalty of 3.0 and a length penalty of 0.1 was used throughout. The sequences of 64 non-homologous structures were used to generate the helix-turn 8 profile, 51 for HT12, 36 for TH8 and 35 for TH12.

Table 4. Consensus sequence of each profile with most likely amino acids at each residue position[a]

	Residue number											
	1	2	3	4	5	6	7	8	9	10	11	12
HT8	hpl[b]	L	m,l	k	hpl	k	G	m				
HT12	e	A	a	hpb[c]	L	k,q	hpl	hpb	G	.[d]	x[e]	V
TH8	L	S	e,d	S,G	B,D,N	y	K	S				
TH12	hpl	.	T	A	E,D	V	a	A	A	L,M	k,q	K

[a] A capital letter (one letter amino acid code) signifies a weighting factor of ⩾ 0.5; lowercase is weighting ⩾ 0.3 and < 0.5.
[b] hpl—hydrophilic amino acids.
[c] hpb—hydrophobic amino acids.
[d] .—no amino acids had a weighting factor ⩾ 0.3.
[e] The amino acid set a, b, d, e, t g, k, p, s, t all had a 0.3 weighting.

The PIR database of amino acid sequences was scanned for sequences which had a close alignment to that of each sequence profile. The alignment of the profile to an amino acid sequence was scored on the basis of the Dayhoff evolutionary metric matrix with a penalty factor for each gap [37].

One restriction of the PROFILESEGMENT program, as currently implemented, is that only the "best" alignment found for each protein is reported. Consequently, the procedure does not report multiple occurrences of a close alignment to the profile in one protein. Table 5 shows the alignment scores of each profile to the database. The score for TH12 was significantly better for the best 100 hits to the PIR database versus the entire database. This was due to a single segment of hemoglobin as identified by the TH12 profile. Since there are more than 100 variants of hemoglobin in the PIR database this search score was artificially high.

Table 5. Profile search of amino acid sequence databases

Template	Maximum score[a]	Protein Identification Resource Database		
		All entries[b]	Top 100[c]	Brookhaven database[d]
Helix-turn 8	3.30	2.31±0.30	2.87±0.08	2.33±0.28
Helix-turn 12	5.10	3.26±0.44	4.02±0.59	3.36±0.35
Turn-helix 8	4.70	3.04±0.42	3.78±0.70	3.10±0.37
Turn-helix 12	6.20	3.84±0.62	5.54±0.07	4.02±0.63

[a] Score is based upon alignment metric matrix of the number of conserved residues less a penalty for introduced gaps.
[b] Average score of all 6862 sequences in release 15.0 of the PIR database.
[c] Average score for the 100 sequences which matched closest to the profile.
[d] Average score for the 82 sequences which are the non-homologous sequences corresponding to known structures in the Brookhaven database of better than 2.5 Å resolution.

The ability of the profiles to correctly identify structural elements in amino acid sequences is summarized in table 6. The 12 residue templates had, on average, a higher discriminatory power than the 8 residue templates. In neither case were the profiles useful for predictive purposes. The number of sequences which were incorrectly identified as the "best" hit by PROFILEGAP was high at some 50%. Since only one hit is reported it is uncertain if any of the segments classified under "Multiple" in table 6 could be correctly identified by this procedure.

Table 6. Distribution of the "best" hits found by each profile sequence[a]

	Number of sequences found			
	Helix-turn 8	Helix-turn 12	Turn-helix 8	Turn-helix 12
Found	5 (7.8%)	6 (11.7%)	4 (11.1%)	6 (17.1%)
Missed	32 (50.0%)	21 (41.2%)	16 (44.4%)	18 (51.4%)
Multiple[b]	27 (42.2%)	24 (47.1%)	16 (44.4%)	11 (44.4%)

[a] Checked against a database of 82 unique sequences which relate to the non-homologous entries in the Brookhaven database of resolution <2.5 Å.

[b] If multiple entries of a structural element exist within a protein only the best hit is reported by PROFILEGAP. The number of extra entries which could not be found are listed as "Multiple".

4. Discussion

The ability of a given protein sequence to rapidly and reproducibly adopt a single major backbone fold is believed to be inherent to its linear amino acid code. However, the initial sequence-specific signals which are associated with the initiation of the folding process are still unknown. Routes or pathways of folding have been proposed for a number of proteins [13]. Certain sites (e.g., certain turns stabilized by a few hydrogen bonds) have a higher degree of structural compactness and may be the primary cores at which folding was originated. The events associated with subsequent side-chain/side-chain stabilizations and further main-chain hydrogen bonds are only open to speculation at this point.

To make the transition between a novel linear amino acid sequence and a three-dimensional structure the protein modeler will need to be able to identify the critical sites necessary for the determination of the overall fold of the protein. This requires, however, the availability of coordinate sets for compact structures and the range of amino acids which can be used to create these sequences.

It is difficult, at this time, to assign structural elements from a protein to an average coordinate template from a family of possibilities. In this work, a rather arbitrary cutoff of a high rmsd of main-chain atoms was chosen. This may not be a very sensitive indicator of structural homology. Application of cluster analysis to side-chain atom contact plots, or to side-chain rmsd values, along with solvent accessibility values at each residue may be useful to help further categorize the fragments and thus better define the template [38]. The accuracy of the turn-helix 8 template in the turn region as compared to the relative diffuseness at the turn region of the turn-helix 12 template illustrates this point well. Also, template definition may be improved during the superimposition procedure. In this work a rigid body rotation/translation algorithm was applied. An alternative would be to use a dynamic algorithm which could allow for breaks in the backbone chain during superimposition [39]. This will be of particular importance for the preparation of larger domain templates.

Once a particular structural template has been defined all sequences which give rise to it can be readily identified. The variability of the amino acids at each residue position over the template region is known as its sequence profile. These profiles are dependent upon the correct sequence alignment among many proteins. Obviously, knowledge of the structure is the ultimate check of the sequence alignment. Application of the standard Needleman-Wunsch algorithm to a small number of sequences will continue to suffer from the well-known alignment problem in which residues that occupy the same three-dimensional volume are often not equated. As a rule of thumb, if the structure is unknown but some 20+ homologous sequences are known, the correct alignment can probably be achieved.

In the absence of structure, a diagnostic sequence profile can still be prepared for certain elements. For example, the consensus profile for the DNA binding zinc finger motif has been defined [13,40].

The metric matrix of Dayhoff (based upon evolutionary relationships) which is used during the sequence alignment procedure may not be appropriate in all cases. It has been shown, in certain structural elements, that otherwise conservative replacements are not possible. For example, the replacement of aspartic acid by glutamic acid is not possible at the N-cap position of an α helix [16].

The identification, preparation, and application of these profiles is still a matter of some debate [41]. For example, if the domain of interest is large, as in the case of a globin fold, it is a reasonably straightforward matter to achieve a correct sequence alignment among many homologous sequences. To be useful for the modeling of proteins *de novo*, significantly shorter domains or substructural elements must be accurately identified: the profile sequences of elements such as α helices or β turns may not be sufficiently specific to discriminate their existence in a sequence. The procedure may thus be limited to finding only a few very specific substructural elements or large folded domains.

If a specific element or fold has been identified from a given structure, a statistically large sample of sequences relating to the template will be required to show the range of residues which can occupy any particular site. The databases of structure and sequences may still be too small to allow for statistical certainty at this time [41].

In the next stage of model building the zones of known structure are joined together to create a range of folding possibilities [42,43]. All the residues are set to alanine except for glycine and proline: this restricts the number of degrees of freedom in the folding problem. Distance geometry or combinatorial approaches can be used to fold the backbone [44]. This is a severely underdetermined system and additional information is certainly needed to constrain the system. The principal restrictions used to restrain the system can be understood easily enough: no atomic overlap; residues should be closely packed; hydrogen bonds are often formed [45]; charged residues are most often found on the surface [46]; restricted conformational possibilities for disulfide bonds [47] and proline residues [48]; sequence dependant statistical data [49,50] such as (flexibility, hydrophilicity, surface accessibility); side-chain volumes; average number of contacts for residues in given substructural regions [36]; Ramachandran plot preferences for phi, psi angles; and any known biochemical information such as disulfide bonding patterns, or specific residues which come together to form an active site.

A major assumption of this approach is that interactions between defined sub-structural domains will affect primarily the details of the side-chain packings [51]: the backbone configuration will remain relatively constant during subsequent model building steps. The placement of side-chains *de novo* is clearly a very difficult job. However, various models have hand-built the core of a protein with surprising ease [52,53]. The methodology to discriminate between competing core packing motifs is still under development. This level of precision, in the preparation of models, is beyond the scope of this work.

These models will be of interest from a variety of standpoints. First, by comparing the variety of ways of joining structural fragments it may be possible to identify why certain motifs are favoured in nature. That is, certain amino acids at specific points may lead to one particular fold. This can be seen most clearly with the role of glycine in allowing certain turn types to exist. Also, the refinement of x-ray crystal structures can also benefit from this approach. A current version of the graphics program FRODO incorporates a library of fragments which can be laid into the electron density map and thus help speed the process of interpretation and refinement [54].

A library of average secondary and super-secondary templates and their associated sequence profiles is currently in preparation. Due to the small size of the databases, the discriminatory power of these profiles may be low. However, the average coordinate sets will still be very useful for general modeling purposes.

5. Acknowledgments

The author thanks Dr. Shoshana Wodak for the preprint and Drs. Steve Bryant, Bob Bruccolerri, and John Moult for helpful discussions.

6. References

[1] Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thorton, J. M., Knowledge-based prediction of protein structures and the design of novel molecules, Nature **326**, 247 (1987).

[2] Moult, J., and James, M. N. G., An algorithm for determining the conformation of polypeptide segments by systematic search, Proteins: Structure, Function and Genetics **1**, 146 (1986).

[3] Dill, K., Protein surgery, Protein Eng. **1**, 369 (1987).

[4] Jones, T. A., and Thirup, S., Using known substructures in protein model building and crystallography, EMBO J. **5**, 819 (1986).

[5] Snow, M. E., and Amzel, L. M., Calculating three-dimensional changes in protein structure due to amino-acid substitutions: The variable region of immunoglobulins, Proteins: Structure, Function and Genetics **1**, 267 (1986).

[6] Summers, N. L., Carlson, W. D., and Karplus, M., Analysis of side-chain orientations in homologous proteins, J. Mol. Biol. **196**, 175 (1987).

[7] Bruccoleri, R. E., and Karplus, M., Prediction of the folding of short polypeptide segments by uniform conformational sampling, Biopolymers **26**, 137 (1987).
[8] Blundell, T. L., Bedarkar, S., and Humbel, R. E., Tertiary structures, receptor binding, and antigenicity of insulin like growth factors, Fed. Proc., Fed. Am. Soc. Exp. Biol. **42**, 2592 (1983).
[9] Heckel, A., and Hasselbach, K. M., Prediction of the three-dimensional structure of the enzymatic domain of t-Pa, J. Comp. Aided Molec. Design **2**, 7 (1988).
[10] Chothia, C., Lesk, A. M., Levitt, M., Amit, A. G., Mariuzza, R. A., Phillips, S. E. V., and Poljak, R. J., The predicted structure of immunoglobulin D 1.3 and its comparison with the crystal structure, Science **233**, 755 (1986).
[11] Yada, R. Y., Jackman, R. L., and Nakai, S., Secondary structure prediction and determination of proteins—a review, Int. J. Peptide Protein Res. **31**, 98 (1985).
[12] Kabsch, W., and Sander, C., How good are predictions of protein secondary structure?, FEBS Lett. **155**, 179 (1983).
[13] Gribskov, M., Hemyak, M., Edenfield, J., and Eisenberg, D., Profile scanning for three-dimensional structural patterns in protein sequences, CABIOS **4**, 61 (1988).
[14] Taylor, W. R., Pattern matching methods in protein sequence comparison and structure prediction, Protein Eng. **2**, 77 (1988).
[15] Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., and Fletterick, R. J., Turn prediction in proteins using a pattern-matching approach, Biochemistry **25**, 266 (1986).
[16] Richardson, J. S., and Richardson, D. C., Amino acid preferences for specific locations at the ends of α helices, Science **240**, 1648 (1988).
[17] Barlow, D. J., and Thornton, J. M., Helix geometry in proteins, J. Mol. Biol. **201**, 601 (1988).
[18] Bashford, D., Chothia, C., and Lesk, A. M., Determinants of a protein fold, J. Mol. Biol. **196**, 199 (1987).
[19] Barton, G. J., and Sternberg, M. J. E., A strategy for the rapid multiple alignment of protein sequences, J. Mol. Biol. **198**, 327 (1987).
[20] Schiff, C., Corbet, S., and Fougereau, M., The Ig germline gene repertoire: economy or wastage?, Immunology Today **9**, 10 (1988).
[21] Chothia, C., and Lesk, A. M., The relation between the divergence of sequence and structure in proteins, EMBO J. **5**, 823 (1986).
[22] Sternberg, M. J. E., and Thornton, J. M., Prediction of protein structure from amino acid sequence, Nature **271**, 15 (1978).
[23] Ponder, J. W., and Richards, F. M., Tertiary templates for proteins, J. Mol. Biol. **193**, 775 (1987).
[24] Go, M., Modular structural units, exons and function in chicken lysozyme, Proc. Natl. Acad. Sci. USA **80**, 1964 (1983).
[25] Crippen, G., The tree structural organization of proteins, J. Mol. Biol. **126**, 315 (1978).
[26] Richards, F. M., and Kundrot, C. E., Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure, Proteins: Structure, Function and Genetics **3**, 71 (1988).
[27] Wodak, S. J., and Janin, J., Location of structural domains in proteins, Biochemistry **20**, 6544 (1981).
[28] Rashin, A. A., Location of domains in globular proteins, Nature **291**, 85 (1981).
[29] Rose, G. D., Hierarchic organization of domains in globular proteins, J. Mol. Biol. **134**, 447 (1979).
[30] Lesk, A. M., and Rose, G. D., Folding units in globular proteins, Proc. Natl. Acad. Sci. USA **78**, 4304 (1981).
[31] Zefus, M. H., Continuous compact protein domains, Proteins: Structure, Function and Genetics **2**, 90 (1987).
[32] Plochocka, D., Zielenkiewicz, P., and Rabczenko, A., Hydrophobic microdomains as structural invariant regions in proteins, Protein Eng. **2**, 115 (1988).
[33] Bernstein, F. C., Koetzle, T. G., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M., The protein databank: A computer based archival file for macromolecular structure, J. Mol. Biol. **122**, 535 (1977).
[34] Devereux, J., Haeberli, P., and Smithies, O., A comprehensive set of sequence analysis programs for the VAX, Nucl. Acid. Res. **12**, 387 (1984).
[35] George, D. G., Barker, W. C., and Hunt, L. T., The protein identification resource (PIR), Nucl. Acid. Res. **14**, 11 (1986).
[36] Reid, L. S., and Thornton, J. M., in Protein Structure, Folding and Design 2, Alan R. Liss, Inc. (1987) pp. 92–102.
[37] Dayhoff, M. O., (ed.) Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, DC, Vol. 5, Suppl. 3 (1978).
[38] Samorjai, R., personal communiation (1988).
[39] Zucker, M., personal communication (1988).
[40] Reid, L. S., manuscript in preparation.
[41] Rooman, M. J., and Wodak, S. J., Reasons underlying low success score of protein structure predictions, Nature, in press (1988).
[42] Ptitsyn, O. B., Random sequences and protein folding, J. Molec. Struc. (Theochem) **123**, 45 (1985).
[43] Goel, N. S., Rouyanian, B., and Sanati, M., On the computation of the tertiary structure of globular proteins. III Inter-residue distances and computed structures, J. Theor. Biol. **99**, 705 (1982).
[44] Cohen, F., and Kuntz, I. D., Prediction of the three-dimensional structure of human growth hormone, Proteins: Structure, Function and Genetics **2**, 162 (1987).
[45] Baker, E. N., and Hubbard, R. E., Hydrogen bonding in globular proteins, Prog. Biophys. Molec. Biol. **44**, 97 (1984).
[46] Lawrence, C., Auger, I., and Mannella, C., Distribution of accessible surfaces of amino acids in globular proteins, Proteins: Structure, Function and Genetics **2**, 153 (1987).
[47] Thornton, J. M., Disulphide bridges in globular proteins, J. Mol. Biol. **151**, 261 (1981).
[48] Chothia, C., Principles that determine the structure of proteins, Annu. Rev. Biochem. **53**, 537 (1984).
[49] Bryant, S. H., and Amzel, L. M., Correctly folded proteins make twice as many hydrophobic contacts, Int. J. Peptide Protein Res. **29**, 46 (1986).
[50] Jameson, B. A., and Wolf, H., The antigenic index: A novel algorithm for predicting antigenic determinants, CABIOS **4**, 181 (1988).
[51] Narayana, S. V. L., and Argos, P., Residue contacts in protein structures and implications for protein folding, Int. J. Peptide Protein Res. **24**, 25 (1984).
[52] Moult, J., personal communication (1988).
[53] Reid, L. S., and Thornton, J. M., Rebuilding flavodoxin from Cα coordinates—a test study, Proteins: Structure, Function and Genetics, submitted, (1988).
[54] Jones, A. T., and Thirup, S., Using known substructures in protein model building and crystallography, EMBO J. **5**, 819 (1986).

Appendix 1. Stage II Template Coordinates with an Average Standard Deviation Derived from the Coordinates Used to Create the Template

Helix-turn 8

Atom No.	Atom Type	Residue No.	X	Y	Z	Std dev
1	N	1	−0.231	−1.983	6.100	0.3536
2	CA	1	−1.279	−2.473	5.293	0.3663
3	C	1	−1.670	−1.514	4.204	0.2730
4	O	1	−1.878	−1.884	3.090	0.2863
5	N	2	−1.715	−0.263	4.546	0.2890
6	CA	2	−2.045	0.777	3.600	0.3193
7	C	2	−0.999	0.886	2.546	0.2400
8	O	2	−1.331	1.030	1.389	0.9320
9	N	3	0.229	0.799	2.934	0.2600
10	CA	3	1.304	0.896	1.998	0.3640
11	C	3	1.288	−0.262	1.026	0.3410
12	O	3	1.570	−0.097	−0.127	0.4650
13	N	4	0.939	−1.408	1.507	0.2603
14	CA	4	0.883	−2.585	0.691	0.3250
15	C	4	−0.287	−2.530	−0.268	0.3126
16	O	4	−0.161	−2.870	−1.405	0.4597
17	N	5	−1.399	−2.123	0.189	0.2237
18	CA	5	−2.603	−2.098	−0.605	0.3057
19	C	5	−2.655	−1.002	−1.620	0.2207
20	O	5	−3.177	−1.174	−2.674	0.3367
21	N	6	−2.130	0.097	−1.302	0.2103
22	CA	6	−2.170	1.238	−2.173	0.8410
23	C	6	−0.962	1.446	−2.931	0.3200
24	O	6	−0.842	2.099	−3.812	0.4777
25	N	7	−0.067	0.907	−2.602	0.5613
26	CA	7	1.102	1.030	−3.268	0.7877
27	C	7	2.119	1.717	−3.210	0.4600
28	O	7	2.510	2.168	−3.743	0.8210
29	N	8	2.535	1.783	−2.557	0.8350
30	CA	8	3.534	2.396	−2.419	0.6900
31	C	8	4.547	2.529	−2.273	0.6777
32	O	8	5.041	2.469	−2.123	0.9867

Helix-turn 12

Atom No.	Atom Type	Residue No.	X	Y	Z	Std dev
1	N	1	6.771	3.843	−3.190	0.4603
2	CA	1	6.583	2.570	−2.619	0.4447
3	C	1	5.401	2.531	−1.748	0.3917
4	O	1	4.618	1.588	−1.770	0.4111
5	N	2	5.256	3.543	−1.006	0.4167
6	CA	2	4.159	3.630	−0.128	0.4980
7	C	2	2.846	3.630	−0.836	0.4110
8	O	2	1.891	2.995	−0.433	0.4817
9	N	3	2.810	4.300	−1.894	0.3430
10	CA	3	1.619	4.364	−2.674	0.4053
11	C	3	1.201	3.041	−3.200	0.3360
12	O	3	0.036	2.665	−3.225	0.4160
13	N	4	2.151	2.335	−3.606	0.3030
14	CA	4	1.915	1.028	−4.114	0.3960
15	C	4	1.370	0.129	−3.092	0.3000
16	O	4	0.436	−0.637	−3.316	0.3800
17	N	5	1.937	0.232	−1.976	0.3350
18	CA	5	1.494	−0.577	−0.918	0.5107
19	C	5	0.098	−0.310	−0.534	0.4610
20	O	5	−0.706	−1.198	−0.297	0.5497
21	N	6	−0.211	0.905	−0.536	0.4593
22	CA	6	−1.528	1.305	−0.216	0.5830
23	C	6	−2.545	0.807	−1.186	0.4630
24	O	6	−3.641	0.399	−0.837	0.5450
25	N	7	−2.180	0.818	−2.381	0.3763
26	CA	7	−3.062	0.384	−3.413	0.4400
27	C	7	−3.411	−1.052	−3.342	0.3443
28	O	7	−4.461	−1.475	−3.681	0.5333
29	N	8	−2.561	−1.782	−2.878	0.2570
30	CA	8	−2.779	−3.177	−2.729	0.3543
31	C	8	−3.386	−3.530	−1.452	0.3693
32	O	8	−3.820	−4.442	−1.241	0.6383
33	N	9	−3.419	−2.846	−0.625	0.5010
34	CA	9	−3.988	−3.092	0.602	0.6870
35	C	9	−3.472	−3.302	1.694	0.4170
36	O	9	−3.852	−3.763	2.527	0.6613
37	N	10	−2.597	−2.934	1.708	0.5007
38	CA	10	−2.020	−3.047	2.731	0.8130
39	C	10	−1.676	−2.232	3.735	0.6440
40	O	10	−1.700	−1.500	3.789	1.0103
41	N	11	−1.407	−2.366	4.529	0.6257
42	CA	11	−1.084	−1.632	5.543	0.7487
43	C	11	−0.016	−1.078	5.836	0.7463
44	O	11	0.371	−1.087	5.826	1.1897
45	N	12	0.449	−0.618	6.098	0.8347
46	CA	12	1.486	−0.070	6.444	1.1203
47	C	12	2.243	0.245	6.887	0.9610
48	O	12	2.383	0.463	7.146	1.1600

Turn-helix 8

Atom No.	Atom Type	Residue No.	X	Y	Z	Std dev
1	N	1	3.667	0.616	6.610	0.3767
2	CA	1	3.517	0.284	5.297	0.2450
3	C	1	3.322	1.484	4.418	0.4580
4	O	1	2.666	2.419	4.814	0.2460
5	N	2	3.860	1.427	3.246	0.1547
6	CA	2	3.676	2.481	2.262	0.1453
7	C	2	2.261	2.433	1.709	0.1257
8	O	2	1.623	1.370	1.672	0.1637
9	N	3	1.771	3.575	1.305	0.1207
10	CA	3	0.443	3.688	0.710	0.1360
11	C	3	0.281	2.769	−0.484	0.1067
12	O	3	−0.790	2.179	−0.670	0.1417
13	N	4	1.330	2.632	−1.261	0.1073
14	CA	4	1.327	1.777	−2.427	0.1503
15	C	4	1.094	0.326	−2.074	0.1470
16	O	4	0.347	−0.367	−2.754	0.1967
17	N	5	1.687	−0.119	−0.996	0.1573
18	CA	5	1.523	−1.484	−0.538	0.2087
19	C	5	0.120	−1.715	−0.035	0.2083
20	O	5	−0.442	−2.786	−0.229	0.2430
21	N	6	−0.406	−0.711	0.601	0.1953
22	CA	6	−1.758	−0.803	1.105	0.2440
23	C	6	−2.778	−0.889	−0.008	0.2023
24	O	6	−3.750	−1.648	0.070	0.2610
25	N	7	−2.539	−0.133	−1.032	0.2057
26	CA	7	−3.424	−0.139	−2.184	0.2483
27	C	7	−3.393	−1.456	−2.907	0.2007
28	O	7	−4.418	−1.928	−3.388	0.2500
29	N	8	−2.253	−2.063	−2.938	0.2070
30	CA	8	−2.109	−3.359	−3.553	0.2777
31	C	8	−2.861	−4.421	−2.817	0.2540
32	O	8	−3.486	−5.278	−3.413	0.3157

Turn-helix 12

Atom No.	Atom Type	Residue No.	X	Y	Z	Std dev
1	N	1	−0.735	5.059	7.604	1.1313
2	CA	1	−0.748	4.906	7.024	0.8767
3	C	1	−1.257	4.780	5.974	0.6573
4	O	1	−1.255	4.475	5.638	0.9323
5	N	2	−1.683	5.016	5.458	0.8573
6	CA	2	−2.198	4.889	4.410	1.003
7	C	2	−2.401	5.125	3.289	0.5843
8	O	2	−2.793	5.004	2.870	0.9253
9	N	3	−2.111	5.432	2.839	0.4377
10	CA	3	−2.258	5.666	1.791	0.549
11	C	3	−1.915	4.971	0.577	0.3917
12	O	3	−1.901	3.838	0.498	0.573
13	N	4	−1.619	5.645	−0.365	0.2827
14	CA	4	−1.259	5.093	−1.611	0.2787
15	C	4	−0.082	4.174	−1.563	0.2807
16	O	4	−0.090	3.125	−2.116	0.3837
17	N	5	0.916	4.537	−0.889	0.358
18	CA	5	2.096	3.735	−0.759	0.4317
19	C	5	1.864	2.431	−0.036	0.434
20	O	5	2.366	1.388	−0.428	0.4293
21	N	6	1.127	2.492	1.011	0.4647
22	CA	6	0.826	1.310	1.771	0.492
23	C	6	−0.029	0.358	0.992	0.3647
24	O	6	0.181	−0.856	1.042	0.391
25	N	7	−0.960	0.905	0.272	0.3097
26	CA	7	−1.816	0.097	−0.542	0.309
27	C	7	−1.036	−0.670	−1.554	0.2067
28	O	7	−1.271	−1.844	−1.782	0.2703
29	N	8	−0.108	0.001	−2.161	0.1563
30	CA	8	0.720	−0.614	−3.156	0.232
31	C	8	1.547	−1.747	−2.587	0.2053
32	O	8	1.678	−2.792	−3.187	0.2853
33	N	9	2.069	−1.532	−1.443	0.2403
34	CA	9	2.873	−2.525	−0.774	0.324
35	C	9	2.061	−3.765	−0.411	0.27
36	O	9	2.502	−4.896	−0.620	0.2783
37	N	10	0.915	−3.533	0.088	0.278
38	CA	10	0.039	−4.627	0.457	0.3543
39	C	10	−0.393	−5.429	−0.719	0.2913
40	O	10	−0.458	−6.652	−0.669	0.3537
41	N	11	−0.686	−4.748	−1.776	0.2387
42	CA	11	−1.093	−5.406	−2.971	0.3363
43	C	11	−0.027	−6.302	−3.511	0.2957
44	O	11	−0.287	−7.416	−3.939	0.3897
45	N	12	1.158	−5.818	−3.454	0.2563
46	CA	12	2.280	−6.571	−3.900	0.368
47	C	12	2.481	−7.825	−3.088	0.341
48	O	12	2.767	−8.885	−3.593	0.462

Appendix 2. Sequence Profiles for Each Template

Helix-turn 8

Residue No.	Type[a]	Amino acid A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
1	E	0.4	0.3	−0.1	0.4	0.4	−0.3	0.3	0.1	0.2	0.1	0.1	0.1	0.2	0.1	0.4	0.0	0.2	0.1	0.2	−0.5	0.1	−0.2	0.4
2	L	0.2	−0.1	−0.2	−0.1	−0.1	0.4	0.0	−0.1	0.4	−0.1	0.6	0.5	−0.1	−0.1	0.0	−0.1	0.1	0.1	0.4	0.0	0.1	0.1	−0.1
3	L	0.1	0.1	−0.2	0.0	0.1	0.1	0.0	0.2	0.2	0.0	0.3	0.3	0.1	0.0	0.2	0.1	0.0	0.1	0.2	−0.1	0.1	0.0	0.1
4	K	0.2	0.1	−0.1	0.1	0.2	−0.1	0.1	0.1	0.2	0.3	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.2	0.2	−0.2	0.1	−0.1	0.2
5	E	0.3	0.3	0.0	0.3	0.3	−0.2	0.2	0.2	0.0	0.2	−0.1	0.0	0.3	0.1	0.2	0.1	0.3	0.2	0.0	−0.2	0.1	−0.1	0.3
6	K	0.2	0.2	−0.1	0.1	0.1	−0.1	0.1	0.1	0.1	0.3	0.1	0.2	0.2	0.1	0.2	0.1	0.2	0.1	0.1	0.0	0.1	−0.1	0.1
7	G	0.4	0.5	−0.1	0.5	0.4	−0.3	0.8	0.1	−0.1	0.0	−0.2	0.0	0.4	0.2	0.3	−0.1	0.3	0.2	0.2	−0.6	0.1	−0.4	0.3
8	M	0.1	0.1	−0.2	0.0	0.0	0.1	0.0	0.1	0.2	0.2	0.2	0.3	0.1	0.0	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.0	0.1
Total[b]		72	0	14	35	34	29	74	43	32	68	87	31	31	4	41	30	50	24	42	6	10	27	0

[a] This amino acid was identified as the consensus amino acid by profile.
[b] Total number of each amino acid used in the generation of the profile.

Helix-turn 12

Residue No.	Type[a]	Amino acid A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
1	E	0.3	0.2	0.1	0.3	0.4	−0.1	0.3	0.0	0.2	0.1	0.0	0.0	0.2	0.1	0.1	−0.1	0.3	0.3	0.2	−0.5	0.1	−0.2	0.2
2	A	0.5	0.2	0.1	0.2	0.3	−0.2	0.3	0.1	0.2	0.1	0.1	0.1	0.2	0.2	0.2	−0.1	0.2	0.3	0.2	−0.4	0.1	−0.2	0.2
3	A	0.4	0.2	−0.2	0.3	0.3	−0.2	0.2	0.2	0.0	0.1	0.1	0.1	0.3	0.1	0.3	0.1	0.2	0.1	0.1	−0.2	0.1	−0.1	0.3
4	L	0.1	−0.1	−0.1	−0.1	−0.1	0.3	−0.1	0.0	0.4	0.0	0.4	0.4	0.0	−0.1	−0.1	−0.1	0.0	0.1	0.4	0.0	0.1	0.2	−0.1
5	L	0.2	−0.1	−0.1	−0.1	0.0	0.3	−0.1	0.0	0.3	0.0	0.5	0.4	0.0	−0.1	0.0	−0.1	0.0	0.2	0.3	−0.1	0.1	0.1	0.0
6	K	0.1	0.2	−0.4	0.2	0.2	−0.1	0.0	0.2	0.1	0.4	0.2	0.3	0.2	0.0	0.4	0.2	0.0	0.1	0.1	−0.1	0.1	−0.2	0.3
7	E	0.4	0.4	0.0	0.4	0.4	−0.2	0.3	0.1	0.0	0.1	0.0	0.0	0.4	0.1	0.2	−0.1	0.2	0.2	0.0	−0.4	0.1	−0.1	0.3
8	V	0.3	0.0	0.1	0.0	0.0	0.2	0.1	0.0	0.3	0.0	0.3	0.3	0.0	0.1	0.0	−0.1	0.2	0.2	0.3	0.0	0.1	0.0	0.0
9	G	0.4	0.5	0.0	0.6	0.4	−0.5	0.8	0.1	−0.2	0.1	−0.3	−0.1	0.4	0.3	0.4	−0.1	0.3	0.4	0.1	−0.7	0.1	−0.5	0.4
10	A	0.2	0.1	0.2	0.1	0.0	0.1	0.2	0.0	0.1	0.0	0.1	0.1	0.1	0.0	0.0	−0.1	0.2	0.1	0.1	0.0	0.1	0.1	0.0
11	T	0.3	0.3	0.0	0.3	0.3	−0.4	0.3	0.1	0.1	0.3	−0.1	0.0	0.2	0.3	0.2	0.1	0.3	0.3	0.2	−0.4	0.1	−0.3	0.2
12	V	0.3	0.2	0.0	0.2	0.2	−0.1	0.3	0.0	0.3	0.0	0.2	0.2	0.1	0.2	0.1	−0.1	0.2	0.2	0.5	−0.5	0.1	−0.3	0.2
Total[b]		120	0	11	35	74	22	103	17	59	64	98	30	60	25	69	30	70	65	88	12	15	72	1

[a] This amino acid was identified as the consensus amino acid by profile.
[b] Total number of each amino acid used in the generation of the profile.

Turn-helix 8

Residue No.	Type[a]	Amino acid A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
1	L	0.0	−0.2	−0.2	−0.2	0.0	0.6	−0.2	0.0	0.5	−0.1	0.7	0.6	−0.1	−0.1	0.0	−0.2	−0.1	0.0	0.5	0.0	0.1	0.2	0.0
2	S	0.4	0.3	0.4	0.2	0.2	−0.3	0.5	−0.1	0.0	0.2	−0.3	−0.2	0.3	0.4	0.0	0.1	1.0	0.5	0.0	−0.1	0.1	−0.4	0.1
3	D	0.3	0.3	−0.2	0.4	0.4	−0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.3	0.2	0.3	0.0	0.2	0.2	0.1	−0.4	0.1	−0.2	0.3
4	G	0.5	0.4	0.0	0.5	0.5	−0.5	0.6	0.0	−0.1	0.3	−0.3	−0.1	0.3	0.3	0.2	0.0	0.6	0.4	0.0	−0.5	0.1	−0.4	0.4
5	N	0.3	0.6	0.0	0.6	0.5	−0.3	0.4	0.3	−0.1	0.2	−0.2	−0.2	0.6	0.1	0.3	0.0	0.4	0.2	−0.1	−0.3	0.1	−0.2	0.4
6	Y	0.0	−0.3	0.0	−0.4	−0.4	0.4	−0.1	−0.1	0.3	−0.1	0.3	0.1	−0.2	−0.1	−0.2	0.1	0.2	0.0	0.3	−0.1	0.0	0.4	−0.3
7	K	0.2	0.2	−0.3	0.2	0.3	−0.2	0.1	0.1	0.1	0.5	0.1	0.2	0.2	0.0	0.3	0.2	0.1	0.2	0.1	−0.2	0.1	−0.2	0.3
8	S	0.2	0.1	0.1	0.1	0.1	−0.1	0.2	0.0	0.0	0.2	0.0	0.0	0.2	0.1	0.0	0.1	0.5	0.2	0.0	−0.1	0.1	0.0	0.0
Total[b]		42	0	35	45	55	15	22	13	22	59	83	11	35	26	24	5	127	36	27	20	1	25	0

[a] This amino acid was identified as the consensus amino acid by profile.
[b] Total number of each amino acid used in the generation of the profile.

Turn-helix 12

Residue No.	Type[a]	Amino acid																					
		A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
1	E	0.3	0.3	−0.1	0.3	0.3	−0.1	0.3	0.1	0.2	0.1	0.1	0.1	0.2	0.1	0.1	−0.1	0.2	0.2	0.2	−0.4	0.1	−0.1	0.2
2	Y	0.0	0.2	−0.3	0.1	0.1	0.2	0.2	0.1	0.1	−0.1	0.2	0.0	0.2	−0.1	0.0	−0.1	0.1	0.1	0.0	0.0	0.1	0.2	0.0
3	T	0.3	0.4	0.1	0.3	0.2	−0.3	0.5	0.0	0.1	0.2	−0.2	0.0	0.4	0.2	0.1	0.1	0.4	0.6	0.2	−0.4	0.1	−0.3	0.2
4	A	0.6	0.3	0.1	0.4	0.4	−0.5	0.5	0.1	−0.1	0.1	−0.3	−0.2	0.3	0.5	0.3	0.0	0.4	0.3	0.0	−0.8	0.1	−0.3	0.3
5	E	0.5	0.4	−0.2	0.6	0.6	−0.5	0.5	0.2	0.0	0.1	−0.1	−0.1	0.3	0.3	0.4	−0.1	0.2	0.2	0.1	−0.8	0.1	−0.3	0.5
6	V	0.2	0.0	−0.1	0.0	0.0	0.1	0.1	0.1	0.4	−0.1	0.4	0.4	0.0	0.1	0.1	−0.1	0.0	0.2	0.6	−0.3	0.1	0.0	0.0
7	A	0.4	0.2	−0.1	0.2	0.3	−0.1	0.2	0.2	0.1	0.1	0.1	0.2	0.2	0.2	0.3	0.0	0.2	0.2	0.2	−0.2	0.1	−0.2	0.3
8	A	0.8	0.3	0.2	0.3	0.2	−0.3	0.6	−0.1	0.0	0.0	−0.1	0.0	0.3	0.3	0.1	−0.2	0.5	0.3	0.2	−0.4	0.1	−0.3	0.1
9	A	0.5	0.2	0.0	0.3	0.4	−0.2	0.3	0.1	0.2	0.1	0.0	0.1	0.2	0.2	0.2	−0.1	0.3	0.2	0.2	−0.5	0.1	−0.2	0.3
10	L	0.2	−0.1	−0.2	−0.2	−0.1	0.5	−0.2	−0.1	0.4	−0.1	0.6	0.6	−0.1	−0.2	−0.1	−0.1	−0.1	0.1	0.4	0.1	0.1	0.2	−0.1
11	K	0.1	0.3	−0.3	0.3	0.2	−0.2	0.1	0.2	0.1	0.4	0.1	0.2	0.3	0.1	0.4	0.3	0.1	0.1	0.1	−0.1	0.1	−0.3	0.3
12	K	0.2	0.3	−0.2	0.2	0.2	−0.4	0.2	0.1	0.0	0.6	−0.2	0.1	0.2	0.2	0.3	0.5	0.3	0.2	0.0	0.1	0.1	−0.5	0.3
Total[b]		168	0	15	61	98	54	90	33	20	61	91	22	67	39	40	52	67	60	95	14	3	26	0

[a] This amino acid was identified as the consensus amino acid by profile.
[b] Total number of each amino acid used in the generation of the profile.

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

Comparative Modeling of Protein Structure—Progress and Prospects

Volume 94 Number 1 January–February 1989

John Moult

Center for Advanced Research
in Biotechnology
University of Maryland
9600 Gudelsky Drive
Rockville, MD 20850

Comparative modeling of protein structure is a process which determines the three-dimensional structure of protein molecules on the basis of amino acid sequence similarity to experimentally known structures. The procedure is facilitated by the growing database of protein structures obtained from crystallography. In this review a series of stages in the modeling process are identified and discussed. These are: (i) obtaining a reliable amino acid sequence of the structure of interest, (ii) producing a structurally correct sequence alignment, (iii) identifying which structural features are conserved between target and parent structures, (iv) modeling the new pieces of structure, and (v) tests of reliability.

Key words: comparative modeling; electrostatics; hydrophobicity; protein structure; sequence alignment.

Introduction

Soluble, globular protein molecules are now some of the best understood components of biological systems. X-ray structures of several hundred structures [1], together with extensive biochemical studies, have led to detailed models of the mechanism of action of these molecules, particularly enzymes. Ever since it was established that in many cases the amino acid sequence is sufficient to determine the three-dimensional structure of such molecules, without the aid of additional biochemical machinery [2], the task of predicting structure from sequence has received a great deal of attention. However, a solution to the general problem still eludes us. Meanwhile, an accumulation of structures which have been determined by crystallography, together with the realization that most proteins are closely related in sequence to a number of others [3], has opened up the possibility of using database approaches to structure determination. These methods, utilizing the set of known structures, are known as comparative protein modeling, or sometimes homologous modeling. As we shall see, the techniques employed use database information in a very indirect manner, with emphasis usually on numerical algorithms rather than the small amounts of data these draw on.

Usefulness

Comparative modeling is a worthwhile activity in two ways: it tests and extends our knowledge of the principles that determine protein structure, and it allows functional insight to be obtained from sequence data quickly, instead of waiting for the laborious process of protein production, purification, and crystallography. With DNA sequencing technology producing sequences at a very high rate it is likely that many, many proteins will be investigated by modeling rather than by structure determination. As an example of this, consider the case of the T-cell defense mechanism of the immune

system. When a cytotoxic T-cell destroys a foreign cell, a number of genes are activated. Detection of these at the messenger RNA level has led to sequence data. Sequence comparison shows these genes to express serine proteases, facilitating comparative modeling of their three-dimensional structure. The models allow deductions to be made about the specificity of these enzymes [4]. Thus, the picture of the T-cell defense mechanism is enhanced, based on modeling alone.

How Difficult Is It?

The degree of difficulty involved in producing a model of a protein by the methods of comparative modeling depends on two main considerations: How accurate and reliable does the model need to be, and how homologous is its sequence with that of the most closely related known structures? Comparison of how similar related structures are [5], has shown that there is an exponential divergence of structure with decreasing sequence similarity. Fifty percent or more identity of sequence ensures a model for which most regions are accurate to approximately 1 Å at the α-carbon positions. With less than 30% identity of sequence few details of a model can be relied upon. These overall figures are often irrelevant, however. If, for instance, the model is needed for the design of molecules that will bind tightly at some site, then only a local subset of the structure is important. An example of such ligand binding oriented modeling is the popular exercise of modeling of the enzyme renin with a view to designing inhibitors which will act as antihypertensives [6]. In such cases, it matters little if 95% of the model is accurate, if the 5% that is wrong involves the substrate binding site. Further, at least a 1 Å root mean square (rms) accuracy is needed in the relevant regions, for all atoms likely to interact with a ligand. One should be cautious, then, of claims that a model is mostly correct. Modeling the highly homologous regions is often trivial, and their correctness may be as irrelevant as in the case of the curate's egg, that was only rotten in parts.

How to Proceed

Progress in this area has been fairly slow, mainly because of a lack of feedback on the accuracy of the models produced. The situation is now changing, as more previously modeled structures become known through crystallography. In this review, I shall draw on a comparison between modeled and experimentally determined versions of the same protein. The case of the relationship between a trypsin-like molecule from the bacterium Streptomyces gresius (SGT) and bovine trypsin (BT) will serve to illustrate many of the pitfalls that litter the path of model building. SGT is 33% identical in sequence with BT, on a structurally based alignment, and so represents a moderately difficult modeling problem. The structure of SGT was modeled twice [7,8] before the x-ray structure was determined. Read et al. made a careful comparison of the modeling results with the x-ray structure [9]. Armed with such insight, one may identify a set of stages in producing a model, and consider for each of these how reliable current techniques are, and what are the prospects for improving them.

This review is intended as a practical guide to comparative modeling—what to worry about, how to do things, and when to believe your own and other people's claims of success.

Stage 1: Obtaining a Reliable Amino Acid Sequence

If there are errors in the amino acid sequence of the protein to be modeled there must be errors in the resulting structure. Problems in this area can reverberate through the subsequent stages, aggravating already difficult steps. It is common experience for crystallographers to find discrepancies between the amino acid sequence reported in the literature and that indicated by electron density maps. Sometimes these are due to isoforms of the protein, but often they turn out to be sequencing errors. In the case of SGT, the omission of two residues from the amino acid sequence [10] happened to occur in one of the most difficult to align regions, and was one of the reasons why all proposed alignments were wrong (in a structural sense—see below) in that stretch of chain. Useful checks that the modeler should make in this area are to look at the sequences of any related proteins that are available, to compare DNA and amino acid level sequences if possible, and to consult the original sequence literature, if obtainable.

Stage 2: Producing a Structurally-Based Sequence Alignment

In order for modeling to begin, the sequence of the target structure must be aligned with those of the relevant known ones. The alignment required is not necessarily that which produces the greatest

number of identities of residue type between the sequences. Rather, it is the one which correctly assigns structural roles to the residues of the target structure. Comparison of alignments based on sequence with those based on structure shows that once the degree of sequence identity falls below approximately 40%, errors are inevitable. In the case of SGT, sequence alignments with BT range between 78 and 91% accuracy [9] for the structurally equivalent residues. The primary problem here is where to put insertions and deletions in the target structure relative to the known ones.

An important point to bear in mind is that more than one parent structure may be useful as the basis for modeling. For instance, in the case of the Hannuka factor T-cell protease, parts of the molecule are most similar in sequence to rat mast cell protease, while others are closer to bovine trypsin [4]. SGT is generally closest in sequence to BT, but one loop of five residues is similar to an equivalent region which occurs only in chymotrypsin. Greer observed this, and was thus able to correctly determine the structure of that piece [8]. That author has referred to this approach as the "spare parts" approach to comparative modeling.

Techniques are beginning to emerge which hold promise of producing structurally correct alignments. One approach makes use of the observation that insertions and deletions are unlikely to occur in regions of secondary structure, such as helices and sheets, so that a large penalty function can be used in such regions in making the alignment [11].

Consideration of the underlying reason for the structural equivalence of residues leads to a more general approach to this step: An extraordinary property of functional proteins is that each amino acid is immobilized relative to the structure as a whole by interactions with the surrounding residues. In this sense, as Havel has pointed out, they are tensegrity structures. This property distinguishes them from other heterogeneous polymers, and presumably from almost all possible random protein sequences. Thus, any pair of structures with related sequences will to a large degree maintain these stabilizing interactions for each residue, and where this is not the case, new ones must be substituted. Consideration of the conservation of interactions in going from one structure to another provides a potentially powerful method of checking and modifying a sequence alignment.

As an example of how this might work, consider the most problematic region of alignment between SGT and BT:

```
          60                  70                  80
SGT: AAHCYKS- - - - -GIQVRLGEDNINVVEGNEQFI
BT: AAHCVSGSGNNTSITATGGVVDL-QSG-SAVKV
     ---------  --------------------  -------
```

The correct alignment is shown here, with the chymotrypsin residue numbering convention. The bars beneath the sequences indicate residues which are structurally equivalent in the two structures (based on data from [9]). Residues G77 and S79 of SGT are the two not reported in the amino acid sequence. The structures of the central portion of this stretch are quite similar, but there are only two residue identities: I63 and G69. There are alternative alignments possible which also have two identities, always involving alignment of V70 in SGT with V75 of BT. An effective alignment algorithm must therefore be able to choose the correct one of these alternatives. Examination of the BT environment of the residues involved shows how this is possible: I63 is involved in extensive hydrophobic core interactions with residues from three remote parts of the structure, and any alignment which does not position an appropriate hydrophobic residue at this position is clearly not viable. G69 is in a buried tight turn, and the surrounding pieces of chain leave no room for any substantial side chain at this position. Less definitively, it is also in the left-handed alpha helix conformation, reducing the probability that any other residue could be accommodated. In contrast to these severe restrictions, V75 makes no hydrophobic contacts, and a number of different side chains can easily be used instead.

In a full implementation of this method, the contacts of every residue in the target structure will be compared with the contacts for the structurally equivalent residue in the parent structure. Contacts are listed by type: nonpolar to nonpolar, polar to polar, charge to charge, and combinations of these. A scoring scheme is used to assess list similarity. Alternative positions for the insertion or deletion of residues may then be assessed on the basis of the conservation of interaction scores. This approach also has application in the next step in the modeling process.

Stage 3: Identifying Conserved Structural Features

Once a satisfactory sequence alignment has been obtained, a preliminary model can be produced by simply substituting each changed amino acid residue in the parent structure(s) for the appropriate one in the target structure. An assessment must

now be made of which residues have the same relationship to the rest of the structure as in the parent molecule, and which different ones. Traditionally, this step has been done by inspection: An obvious place to begin is with the places where there is good sequence homology between parent and target structure. One would normally assume that such regions will have closely similar structures. Although this is generally true, there are exceptions. In SGT, for example, there is one region of apparent reasonable sequence homology (146–152 [9]) where main chain α-carbon positions differ by more than 1.9 Å compared with the parent BT structure:

```
          140              150
SGT: G W G A N R E - G G S Q Q R Y L
BT: G W G N T K S S G T S Y P D V L
     ----------          -----
```

The method of environment characterization outlined above can be used to appreciate why there is a breakdown in structural similarity in this stretch: G148 is held in position in BT by polar interactions with the side chains of N143, S147, and T149. In SGT, these residues are all atrophied, and nonpolar, becoming A, A, and G, respectively. Interactions of the side chain of S150 in BT are weak, so that it largely relies on the surrounding residues to hold it in position, and these are not conserved. Application of the environment classification algorithm would thus appear to be able to identify which features of structure are conserved in the target molecule.

Stage 4: Building New Structural Regions

At the end of the previous step, a list of regions which need remodeling, ranging from single residue side chains to whole stretches of chain, has been produced. These regions must now be constructed in some manner. A number of approaches are possible:

(a) Human judgment. The most usual method. An operator sitting in front of a graphics system inspects the region to be modeled, and draws on his experience to suggest one or more possible structures. In simple cases, such as positioning a single side chain, this may be effective. In general, though, there are too many possibilities to be considered, and human judgment tends to be more like human prejudice than a rational consideration of the options. In the modeling of SGT, Read et al. [9] found that all attempts to construct segments of three or more residues led to serious errors. There is also a problem of nonreproducibility inherent in such a subjective procedure.

(b) Databases of known structures. Since we are interested in quite short lengths of polypeptide chain, it seems likely that useful information can be obtained directly from the database of known structures. For lengths of chain up to five residues long it is indeed true that the set of known structures will usually contain one or more examples with an rms on α-carbon atoms of 1.0 Å or better to any target structure. The issue becomes one of selection. Sequence homology for such short stretches is not an indication of similarity of conformation, even in the limit of identical sequences [12]. Knowing the positions of the ends of the stretch turns out to be a powerful constraint on possible conformations. For up to five residues, this information may be used to select a small set—typically 5 to 20 conformations, separated from each other by 1 Å rms or more in α-carbon rms space. This is particularly useful for selecting a conformation that best fits a poor quality electron density map [13]. However, when no experimental information is available to limit the choice, the outcome is a set of possible α-carbon traces, and no means of choosing between them. Further, the database is only large enough to provide a description at the α-carbon level—once all atoms of the backbone and side chains are added the number of possible conformations increases dramatically. Such a detailed level of description is needed both to provide a useful model of the structure, and to enable energetic criteria to be used to distinguish between possibilities. Because the number of possible all atom structures is so great, the database of structures will never be large enough to contain all five residue stretch conformations at the 1 Å accuracy level. Still, this method has the merit of speed, and may be used to guide operator thinking to extend the scope of method (a). The graphics program FRODO elegantly incorporates the required database, and other programs will soon do so as well.

(c) Molecular dynamics simulation. The most straightforward of a set of approaches which use energetic criteria in some form to select a conformation. The idea here is that an arbitrary conformation of the region to be modeled is selected, and then a molecular dynamics trajectory of the local region of structure is performed, sufficiently long that the correct conformation will be encountered,

and may be recognized by its low energy compared with the alternatives. Contemporary empirical potentials do seem to be able to represent structure at a level approaching the 1 Å rms level [14], so that from a discrimination standpoint this approach is viable. There are serious difficulties with the length of the simulation required, however. Getting from an arbitrary starting conformation to the correct one may entail rearrangements of the local structure which can only be achieved by unfolding a large part of the protein, so that no affordable simulation will be able to reach the correct structure. Possible approaches to overcoming this obstacle are to use an elevated temperature, making the surmounting of conformational energy barriers more frequent; a lowering of the van der Waals repulsive energies so that pieces of chain may pass through each other; and using a number of starting conformations, in the hope that one of them will be within the convergence range of the method. So far, these approaches have not been demonstrated to be effective.

(d) Distance geometry. If sufficient interactions can be identified, these may be used as restraints to restrict the number of possible conformations, in much the same manner as NOE distances are used to define peptide and protein structure with this method [15]. Possible restraints are connection to the rest of the structure, and liganding to a bound group, such as an ion [16]. It may also be feasible to investigate combinations of possible interactions, such as salt bridges or hydrophobic contacts. Such restrictions may provide a powerful approach to this type of modeling problem, but their potential has yet to be explored.

(e) Systematic Conformational Search (SCS). This method consists of two steps: Generate all possible conformations, and then use some sort of discriminatory functions to choose one close to the correct structure. Several groups [14,17–19] are pursuing this approach, with somewhat different methodologies. Here I will summarize our own work [14].

"All possible conformations" implies sampling the conformational space sufficiently finely that at least one conformation will be generated within the desired deviation from the correct structure. The limit of acceptable deviation is dictated by the ability of the discriminatory functions to identify a structure as nearly correct, and to distinguish which of two possible conformations is actually more correct. In practice, this means sampling the conformational space with a density of about 1 Å rms.

For the discriminatory functions to be effective, an all atom description of the structures is at present required. There are astronomical numbers of such possible structures at the 1 Å rms density level, even for short pieces of chain—of the order of 10^{4n}, where n is the number of residues. Thus the approach is impractical, unless ways can be found of reducing the number of conformations that need to be considered. This can be done by using rules that protein structures obey as filters to reduce the number of conformations. Useful rules are simple: For example, the preference for ranges of dihedral angles, the avoidance of van der Waals clashes, and the restriction of joining the edges of the segments to the rest of the structure. In practice, however, none of these rules is completely obeyed: For instance, the energetic cost of van der Waals clashes is so high that they never occur in real structures. However, when structures are generated by finite sampling, very significant clashes will likely be present in the best structure generated. Such clashes must therefore be allowed for in deciding whether to accept the structure for further consideration. Nevertheless, for lengths of up to seven residues, depending on the sequence, the number of conformations can be reduced to a few thousand, and these may then be evaluated using the discriminatory functions.

Suitable discriminatory functions are the electrostatic energy of each conformation, and the amount of exposed hydrophobic area. Tests have shown that it is possible to first determine the conformation of electrostatically sensitive parts of a structure (main chain, polar and charged side chains) and then to consider the best conformation of hydrophobic side chains. Electrostatic energy is evaluated using an empirical force field [20], together with a solvent reaction field [21]. Exposed hydrophobic area is calculated using a standard surface area algorithm [22], and considering those atoms in the force field which carry zero partial charge to be hydrophobic.

These discriminatory functions are effective. Although they do not always select the very best structure generated, they do find one close to the best, and within approximately 1 Å rms of the correct (x-ray) structure [14]. Provided the previous stages have been successfully carried out, good structural models are produced. Further developments of the method are in progress, and hold promise of extending the number of residues which can be considered.

Stage 5: Evaluation of Reliability

A central difference between modeling and experiment is that in the latter regime it is usually possible to test a conclusion against the data and know how reliable that conclusion is. This is not generally possible with a model. However, there are a number of ways of assessing the reliability of a protein model:

(a) Note any uncertainties in the data for the original sequence, such as differences between amino acid and nucleic acid results, and substitutions compared with related proteins that appear unlikely. Also note uncertain or alternative possible sequence alignments with the parent structure(s).

(b) Evaluate the exposed surface area of different types of residues in the resulting structure, and note any exposed hydrophobes, or buried charges. Although both situations do occur [23,24] they are sufficiently unusual to be worthy of suspicion.

(c) Evaluate the packing of each group in the protein, and note any poorly packed regions, or cavities. Poor packing and cavities are present in proteins [25], but again, their presence in a model serves to focus attention on a particular suspect region.

(d) Evaluate the electrostatic environment of all polar and charged groups. Note any overall unfavorable ones. Examination of refined x-ray structures shows that these are unusual in well ordered regions.

None of these criteria will yield a completely definitive answer. However, they will allow the probability of correctness of any feature in the structure to be assessed. The usefulness of these criteria has recently been reviewed [26]. This information can then be used in deciding whether to accept conclusions concerning function drawn from the model.

References

[1] Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M., J. Mol. Biol. **112**, 535 (1974).
[2] Anfinsen, C. B., Science **181**, 223 (1973).
[3] Baker, W. C., Hunt, L. T., George, D. G., Yeh, L. S., Chen, H. R., Blomquist, M. C., Seibel-Ross, E. T., Elzanowski, A., Hong, M. K., Ferrick, D. A., Blair, J. K., Chen, S. L., and Ledly, R. S., Protein Sequence Database, National Biomedical Research Foundation, Georgetown University Medical Center, Washington, DC.
[4] Murphy, E. P. M., Moult, J., Bleackley, R. C., Gershenfeld, H., Weissman, I. L., and James, M. N. G., Proteins (In press).
[5] Chothia, C., and Lesk, A. M., EMBO J. **5**, 819 (1986).
[6] Blundell, T., Sibanda, B. L., and Pearl, A., Nature **304**, 273 (1983).
[7] Jurasek, L., Olafson, R. W., Johnson, P., and Smillie, L. B., Miami Winter Symp. **11**, 93 (1976).
[8] Greer, J., J. Mol. Biol. **153**, 1027 (1981).
[9] Read, R. J., Brayer, G. D., Jurasek, L., and James, M. N. G., Biochemistry **23**, 6570 (1986).
[10] Olafson, R. W., Jurasek, L., Carpenter, M. R., and Smillie, L. B., Biochemistry **14**, 1168 (1975).
[11] Lesk, A. M., Levitt, M., and Chothia, C., Prot. Eng. **1**, 77 (1986).
[12] Kabsch, W., and Sander, C., Proc. Natl. Acad. Sci. USA **81**, 1075 (1984).
[13] Jones, T. A., and Thirup, S., EMBO J. **5**, 823 (1986).
[14] Moult, J., and James, M. N. G., Proteins **1**, 146 (1986).
[15] Harel, T., Kuntz, I. D., Crippen, G. M., Bull. Math. Biol. **45**, 655 (1983).
[16] O'Neil, K. T., and DeGrado, W. F., Proc. Natl. Acad. Sci. USA **82**, 4954 (1985).
[17] Vasquez, M., and Scheraga, H. A., Biopolymers **24**, 1437 (1985).
[18] Bruccoleri, K. T., and Karplus, M., Biopolymers **26**, 137 (1987).
[19] Fine, R. M., Wang, H., Shenkin, P. S., Yarmush, D. L., and Levinthal, C., Proteins **1**, 342 (1986).
[20] Hagler, A. T., Huler, E. H., and Lifson, S., J. Amer. Chem. Soc. **96**, 6319 (1974); Lifson, S., Hagler, A. T., and Dauber, P., J. Amer. Chem. Soc. **101**, 5111 (1979); and Hagler, A. T., Lifson, S., and Dauber, P., J. Amer. Chem. Soc. **101**, 5122 (1979).
[21] Moult, J., Sussman, F., and James, M. N. G., J. Mol. Biol. **182**, 555 (1985).
[22] Lee, B., and Richards, F. M., J. Mol. Biol. **55**, 379 (1971).
[23] Richards, F. M., Annu. Rev. Biophys. Bioeng. **6**, 151 (1977).
[24] Rashin, A. A., and Honig, B. H., J. Mol. Biol. **173**, 515 (1984).
[25] Rashin, A. A., Iofin, M., and Honig, B., Biochemistry **25**, 3619 (1986).
[26] Novotny, J., Rashin, A. A., and Bruccoleri, R. E., Proteins **4**, 19 (1988).

Volume 94, Number 1, January–February 1989
Journal of Research of the National Institute of Standards and Technology

The Computational Analysis of Protein Structures: Sources, Methods, Systems and Results

Volume 94 Number 1 January–February 1989

Arthur M. Lesk

European Molecular Biology Laboratory, Heidelberg, F.R.G. and MRC Laboratory of Molecular Biology, Cambridge, U.K.

and Anna Tramontano

European Molecular Biology Laboratory, Heidelberg, F.R.G. and International Institute of Genetics and Biophysics, via Marconi, 10, 80215, Naples, Italy

Computational molecular biology is a relatively new specialty that has arisen in response to the very large amount and quality of data currently being produced, including gene and protein sequences ("one-dimensional" information) and nucleic acid and protein structures ("three-dimensional" information). Many important biological investigations can be carried out only through effective computational access to the entire corpus of data. This has stimulated the development of data banks and information retrieval systems. For example, after determination of a new gene sequence, one would like to know whether it is possible to say anything about its structure and function. To try to answer this question one screens the sequence of the corresponding protein for a significant similarity to a protein of known structure. In this article we shall describe the kinds of inferences that are possible if such a relationship is found.

Key words: data banks; molecular biology; molecular graphics; protein structures; structure prediction.

Introduction

The state of information-retrieval systems in molecular biology is currently undergoing rapid change. This is partly a result of the great increase in the sheer amount of data available, and partly the result of advances in computing equipment that have made available very powerful systems capable of supporting high-capacity information storage, demanding calculations, and complex real-time graphics, and a better definition of the roles in the partnership between the program system and the scientist. But it is also the result of our beginning to understand somewhat better the kinds of questions we want to ask. For example, until recently the one-dimensional world of sequence calculations and the three-dimensional world of structure calculations remained aloof from each other; now it is recognized that it is essential to bring both sets of data to bear on problems together. For another example, until recently many people—in the three-dimensional world—would work on a single protein structure or family of structures in isolation. Now we recognize the importance of free and common access to all available proteins, because we can recognize structural themes common to a wide variety of structures.

These three factors—large and rapidly increasing amounts of data, new powerful computer systems, and greater sophistication—are now in collision. Here we shall describe what might emerge.

The Data

Nucleotide sequences contain the blueprints for the development of living organisms. They directly encypher the amino acid sequences of proteins, agents of biological structure and function. Once the amino acid sequence of a protein has been synthesized, it then spontaneously folds to create a unique three-dimensional protein conformation. It is at this point that the linear genetic code is translated into three dimensions.

Nucleic acid sequences, protein sequences, and protein structures are all collected and distributed by data banks.

Nucleic acid sequences are collected by a tripartite association of organizations: GenBank® in the United States of America, with scientists at Los Alamos National Laboratory and Intelligenetics, Inc.; The Nucleotide Sequence Data Bank at the European Molecular Biology Laboratory in Heidelberg, Federal Republic of Germany; and the DNA Data Bank of Japan, at the National Institute of Genetics, in Mishima. These groups collaborate in harvesting data from published journals, and in sharing the results. To an increasing extent, the data banks are receiving data in computer-readable form directly from scientists. The data are converted to standard formats, checked and annotated, and then exchanged among the databanks and distributed to scientists.

It may be interesting to have some standards of comparison for the amounts of data involved. If one base pair is stored as one byte, the genome of the Epstein-Barr virus has 172 kbytes, the genome of the much studied bacterium *E. coli* has 4000 kbytes, the genome of yeast is around 20,000 kbytes, and the human genome is a factor of 1000 above *E. coli* at 4×10^9 bases or 4×10^6 kbytes. The *E. coli* genome stored at 1 byte per base pair has approximately the same number of characters as the Cambridge, England telephone directory. A human genome has about an order of magnitude more characters than the Oxford English Dictionary. The Dictionary in its new printed form is a set of 16 large volumes, and also occupies an entire compact disc.

Protein sequences are collected by another triple partnership. For many years, the group at the National Biomedical Research Foundation in Washington, D.C. maintained the major computer-readable archive of protein sequence data. In addition to collecting, annotating, and distributing sequences, this group has developed a powerful information retrieval system integrated with the data in the Protein Identification Resource (PIR). This group has recently been joined by others in the Federal Republic of Germany and in Japan.

The archive of three-dimensional structures of biological macromolecules is the Protein Data Bank at Brookhaven National Laboratory in New York, U.S.A. It collects the results of structure determinations, primarily by x-ray crystal structure analysis, but with a soupçon of structures determined by neutron diffraction; these to be joined by structures determined by NMR, which has established itself as quite a fruitful source of structural information for relatively small macromolecules. The Crystallographic Data Center in Cambridge, England, maintains a database of small molecular structures determined by x-ray crystallography. This information is extremely useful in studies of the conformations of the component units of biological macromolecules, and for investigations of macromolecule-ligand interactions.

A Task Group of CODATA (Committee on Data of the International Council of Scientific Unions), chaired by Prof. B. Keil of the Institute Pasteur; secretary, Dr. A. Tsugita of the Science University of Tokyo, has been working to try to foster collaboration among data banks, and between the data banks and the scientific community.

Data Distribution

In the past, most of the data banks distributed their contents on magnetic tape or floppy disk, emitting successive releases at standard intervals, typically 3 months apart. Recently there has been some exploration of the use of computer networks for data distribution (as well as for submission of data), using the concept of a Netserver which responds to queries sent in over networks by returning a requested item as a reply (if possible). Other high-density storage media—notably CD-ROM—are also being explored; particularly exciting is the possibility of desk-top information retrieval systems self-contained in a personal computer-CD reader combination.

Information-Retrieval Systems

To say that an understanding of the relationships among the data will provide the keys to breakthroughs in theoretical and experimental biology raises more questions than it answers. The availability of so much data, and the large number

and incomplete definition of the relationships among them, create serious problems of data storage, quality control, and information retrieval. Plans for "mega-projects" such as the sequencing of the entire human and rice genomes will only intensify these challenges.

The rapid increase of computer power in recent years has begun to give us the tools to address these problems, and we must focus on the problems of design of a system to store, check, update, and distribute the incoming data, and then to provide the tools to produce new scientific results. These problems are common to many fields of science. Their general solution, based on an effective data base management system, has advantages that are well known: Applications programmers are relieved from standard management tasks, quality control is easier— redundancy and inconsistency in the data or in its formatting can be reduced, and the integrity of the data thereby more easily maintained.

A molecular biology information system must include both sequence and structural data, and the database management system must be able to answer, directly, most of the questions that investigators want to ask about the relationships among the data. It must also be flexible enough to accommodate new questions. Whether a commercial database management system can be used, or whether modifications or extensive redesign are required, depends on the structure and type of the information, on the manipulations to be performed on the data, and on the universe of user queries. Because the field is evolving with great speed intellectually, it is very hard to foresee the kind of questions we will come up with, even in the next few years. In the design of a database management system for molecular biology it would be fatal to sacrifice flexibility for efficiency.

It is necessary to present structural and nonstructural data in the same framework. Inquiries such as "Is this sequence fragment present in other sequences?" and "Is this structural motif present in other structures?" should be asked by the user within a common framework of dialog. Of course the second type of question is more complicated, for it requires an interface general enough to define a structural motif. Thus the question "Is this new entry similar to any already-existing one?" is relatively well-defined if we are talking about linear (sequence) data, because the answer can be expressed in terms of the number of common residues and standard statistical parameters. In contrast, the same question, in the sense of structural similarity, requires further specification: For example, do we want to know about the overall shape of the molecule, about the relative positions of secondary structure elements, about the geometry of conserved residues in the active site? All these are legitimate inquiries, and such variations place great strain on a query interpreter.

Proteins exhibit both complex topological features and detailed local structural patterns. The careful observation of proteins, one at a time, can help us to define and propose some general principle of protein architecture; the comparative analysis of several structures probe the likelihood that our hypothesis is correct, and devising a general algorithm for testing the hypothesis with the entire available data set can confirm the proposed rule.

A problem facing those who would design a "packaged" system is the difficulty of defining a set of operations that encompasses the needs of the users. We ourselves, after having spent years in analyzing the operations useful in research (trying with only limited success to define a set of "elementary" operations in terms of which most manipulations might be defined) and in constructing software, find it frustrating that when a new project is undertaken it almost always requires the development of new tools.

Thus the database will have to cope with all the problems well-known from traditional "one-dimensional" databases, and also new ones specifically related to a "three-dimensional" database and a subject with widening intellectual horizons. The traditional problems include, for example, the problems of accommodating uncertain and partial data, of updating the system without loss of continuity, and most important of all, the problem of checking the data for consistency. Particular to chemical structural data is the need for a molecular graphics interface. Because features of the database entries must be presented graphically, in a consistent way, the design of the database must include a means of integrating the retrieval of information with molecular graphics packages.

A complete database should provide a flexible graphics interface allowing the user to visualize the atomic details of protein and nucleic acid structures, and some schematic view of their overall shape and secondary and tertiary structure. Whether the interface should include the graphics software or provide a way of interchanging information with the several existing graphics packages is not the most important question. The objection that in the first case the database would be more hardware-dependent will be overcome by the

spread of graphics standards. What is needed in both cases is the availability of a clear definition of objects, representations and views applied to molecular objects.

Molecular Graphics

How do we go about analyzing protein structures? First, we make a general inspection of the structures, using computer graphics. Programs take a set of coordinates and create a visual image on some device. The system gives the user the facility of selecting a portion of the molecule to be shown, selecting the orientation of the picture, and selecting the representation of the structure.

Two basic representations are (1) to show each atom as a sphere, distinguishing different atoms by different colours or shades; (2) to show each bond as a line. The former requires what is called a "raster" device, giving an image with the appearance of a television screen. Typically the image can contain 512×512 or 1024×1024 "pixels," with each point chosen from one of 256 (or more) possible combinations of colour and intensity; thus one might have 16 intensity levels of each of 16 colours.

The second type of representation is called line or vector or calligraphic graphics. Here the technology exists to draw tens of thousands of lines, in different colours, at a refresh speed that does permit real-time rotation, which greatly enhances the observer's perception of the three-dimensional structural relationships. (Real-time rotation of raster pictures is possible in the new generation of graphics workstations, which are now being applied to molecular biology. This capacity has existed for some time, but until recently only in devices of such high cost that they were limited to special applications such as the training of aircraft pilots.) Other important methods of enhancing the perception of three-dimensional structural relationships include stereo, hidden-line removal, and depth cueing (that is, the diminishing of intensity of objects farther from the eyepoint.)

Because of the complexity of protein structures, pictures in which every atom or every bond is shown individually are often uninformative. People have therefore devised simplified or schematic representations. In these, a common grouping of atoms called an α-helix may be shown as a cylinder, and another common grouping of atoms called a strand of β-sheet may be shown as a large arrow (see figs. 1–3).

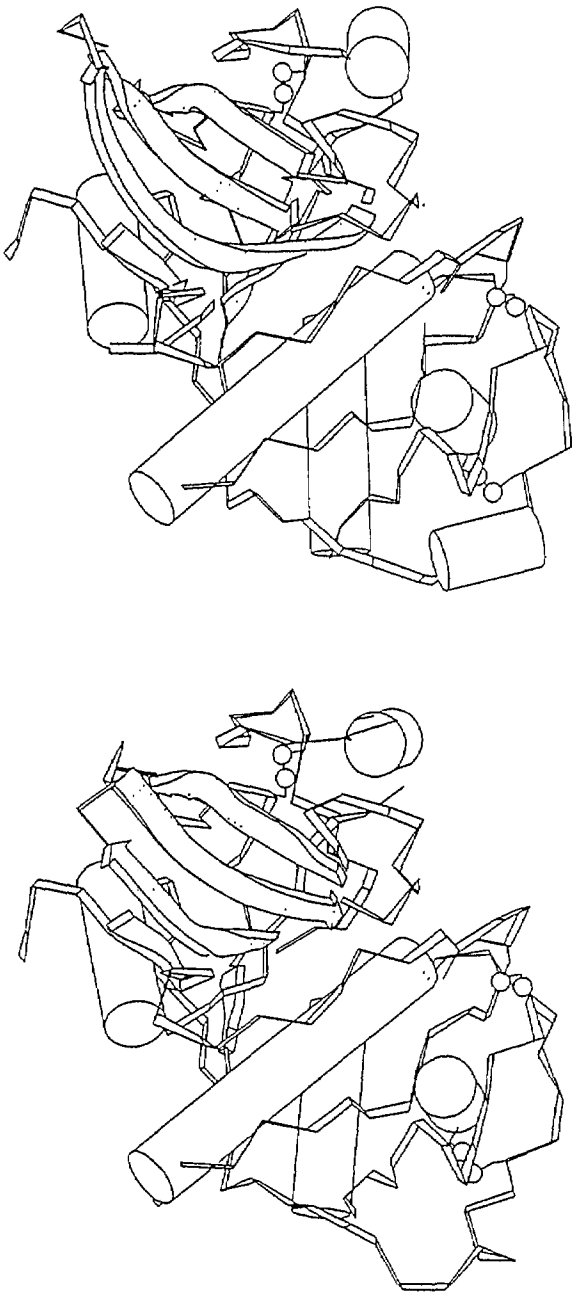


Figure 1. Two closely-related proteins: (a) actinidin [7] and (b) papain [8]. The amino acid sequences of these molecules have about 50% identical residues.

The analysis of a protein into helices, sheets, and other regions (often called loops) is part of the initial investigation of the structure and might be considered analogous to the parsing of a sentence, or at least to the identification of nouns and verbs. Helices and sheets are common arrangements of regions of proteins, stabilized by hydrogen bonds.

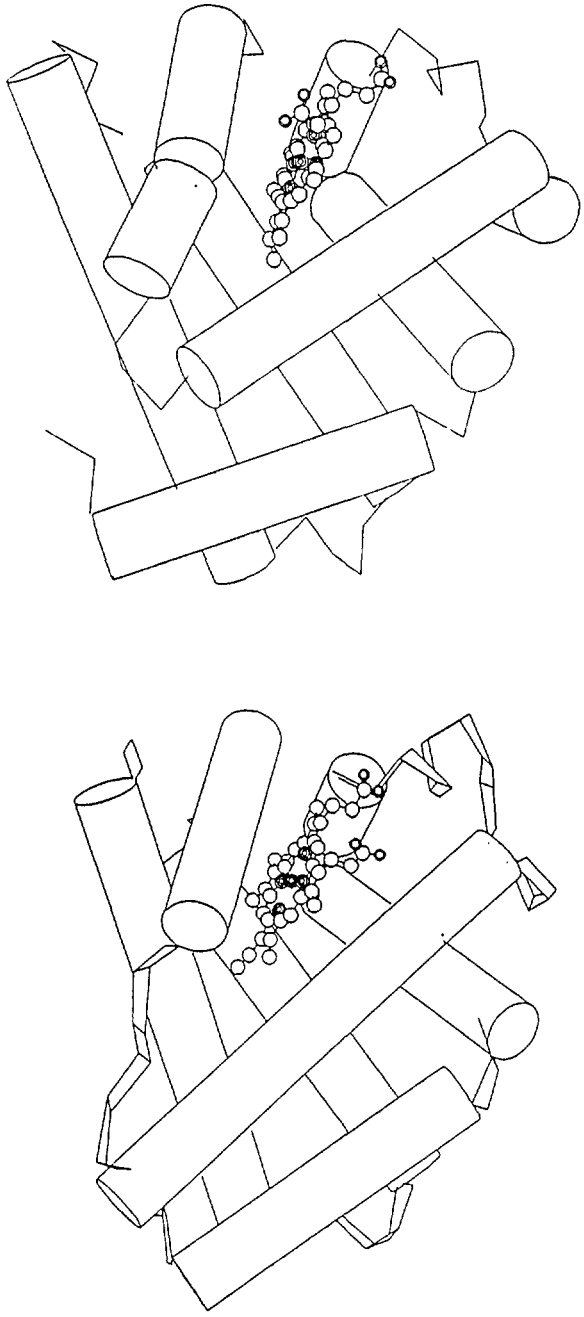


Figure 2. Two quite distantly-related proteins: (a) sperm whale myoglobin [9] and (b) lupin leghaemoglobin [10]. In this case almost the entire chains have the same fold.

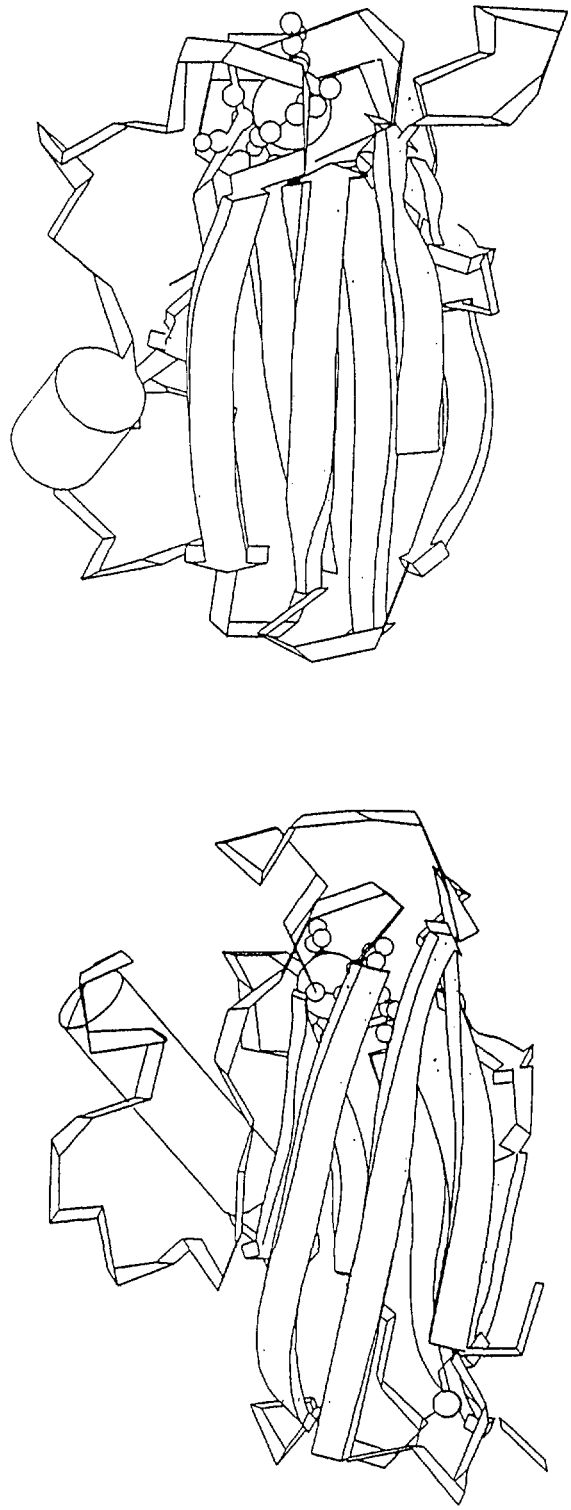


Figure 3. Two other distantly-related proteins: (a) poplar leaf plastocyanin [11] and (b) *A. denitrificans* azurin [12]. In this case the double β-sheet portion of these molecules retains the same fold, but the long loop at the left changes its conformation completely.

They were predicted by Linus Pauling on the basis of physico-chemical principles before the discovery of the first protein structures, myoglobin and haemoglobin, in which the presence of helices was gratifyingly confirmed.

We can identify helices and sheets in proteins either visually, or by the detection of hydrogen bonds by purely numerical analysis of the coordinates, or by geometrical analysis of the positions of the atoms. There exist programs that will take a set of coordinates and produce a set of helix and sheet assignments automatically. Because of the not uncommon "fraying" of the ends of these regular substructures, these programs work fairly well but not perfectly.

Knowing where in the structure the helices and sheets lie, we can create a variety of representations of the structure. Protein structures have been classified into certain basic types on the basis of the types of secondary structures they contain and the spatial relationships between them. Such a diagram will be enough for an expert to place a new structure in the current scheme, or to recognize a real novelty.

Storing images or the coordinates necessary to rebuild them could be transparent to the user if a "molecular graphics metafile" is provided, where a clear definition of the properties of the displayed object is stored. While standardizing the graphic representation of a molecular object is relatively straightforward when dealing with one molecule at a time, several problems arise when more than one molecule has to be displayed in the same coordinate space. Showing two superimposed molecules, or an enzyme together with its substrate, requires in both cases the visualization of two molecules, but the physical meaning of the two double images is completely different. Some operations are allowed in one case (for example in the first case two atoms occupy the same position in space) but forbidden in the other. In other words the "metafile" should also define the possible operations that can be performed in each case, so that the application program, whether it is a part of the database or not, should treat the two cases differently providing the user with the appropriate functions for each.

Protein Modeling

The general ideas presented here can now be illustrated with an important example: the question of modeling the structures of unknown proteins. In order to have a specific framework for this discussion, let us consider a particular problem; one which in fact arises in virtually this exact form.

Suppose we know the structures of two related proteins, for example, the sulphydryl proteases actinidin [7] and papain [8], or the two electron-transport proteins plastocyanin [9] and azurin [10], or sperm whale myoglobin [11] and lupin leghaemoglobin [12] (see figs. 1–3). Suppose someone shows up with a third sequence, of a natural protein of unknown structure related to the other two. What can we say about its structure? (The restriction to natural variants is now important because molecules synthesized in the laboratory have not undergone the trial of natural selection and may not follow the same rules.)

In order to answer this question, we must know how to align the sequences of the known proteins, we must be able to identify and describe the structural differences between the known proteins, and we must be able to know how the differences in the amino acid sequences are related to the structural differences. Deriving this insight from the known proteins we can extrapolate to their unknown relative. Let us consider some of the computational steps we go through, and the nature of the software and hardware that have proved useful.

Let us first dispose of what we might with some temerity call a potential distraction: Someday it may be possible to ignore the fact that the unknown protein is related to others of known structure, and to predict its conformation from physical principles. This is just not possible today (see below).

It follows that, given a new sequence, we must first try to find out whether it is related to proteins of known structure. There are now fairly standard techniques for screening databases of sequences, to pick up many—but not all—relationships. Very distant relationships may elude these procedures, as it is a fact that structural relationships can exist when the overall sequence similarity has diverged so far as to conceal the homology. There are more sensitive methods for picking up members of some individual protein families, by looking for a specific "fingerprint" or "signature" of a protein that may involve only a small fraction of the residues.

If we find that the unknown protein is related to other proteins of known structure, it is possible to draw two conclusions about its conformation:

(1) the structure of the unknown protein is like the structure of the known proteins and

(2) the structure of the unknown protein is unlike the structure of the known proteins.

Although this sounds like something from Alice in Wonderland, both comments are true. The first is a statement that related amino acid sequences determine protein structures that have the same general topology or fold, over at least 50% of the molecule. The second statement points out that amino acid sequence changes produce conformational changes, so that the structure of one of the proteins will be a distorted version of the structure of the other. The extent of the distortion, which limits the quality of the model of the unknown protein that we could build, depends on how far the amino acid sequences have diverged.

We have already discussed how we look at one protein structure at a time. To reason about an unknown protein from related known ones, we must now turn to the question of how we analyze the structural differences between two related proteins.

There is a basic computational tool in comparative structural analysis, which is the geometric superposition of a pair of structures. Given two lists of atoms, which may be regions selected from two proteins, the problem is to find a rotation matrix and translation vector that will optimally superpose the two structures, in a least-squares sense. We must know the proper correspondence of the atoms in the two structures, not a trivial question in the face of insertions and deletions of amino acids in the sequences of proteins.

Fortunately, this is a very simple problem to solve, and several fast and reliable algorithms are available. The result of such a calculation is the optimal geometric transformation, and the root-mean-square (rms) deviation of the atomic positions. It is also possible to list individual atomic deviations and thereby distinguish well-fitting regions from other regions in which structural change has occurred. The operation of performing superpositions of selected regions of proteins is the basic tool of quantitative structural comparison, akin to something as fundamental as pipetting in the laboratory.

What does such analysis tell us about the structural differences between pairs of related proteins? First, it shows that in a family of proteins there is a core of the structure that retains the same basic topology, or fold, and the rest can have a completely different conformation [1,2]. (To explain the idea of the common core of two structures, look at the letters B and R. Considered as structures they have a common core corresponding to the letter P. Outside the common core they differ: at the bottom right B has a loop and R has a diagonal stroke.) In plastocyanin and azurin, the double β-sheet retains its fold but the long loop at a side of the sheet does not. Secondly, it shows that although individual helices and sheets tend to retain their structures fairly rigidly, there are changes in their relative geometrical relationship—shifts and rotations of one relative to another. Using superposition calculations we can measure the magnitude of these shifts and rotations.

What can we then say about the structure of a new protein? The general comment is that the common core that this protein shares with the known structures will have the same fold; but, except in special cases, we cannot predict the structure of the regions outside the core. More specifically, we can relate the fraction of the structure in the core, and the magnitude of the distortions of the core structure, to the divergence of the amino acid sequences. Note that these quantitative results required numerical superposition calculations, not merely looking at the structures. There are numerous program systems that combine interactive graphics with superposition facilities.

The basic rule-of-thumb that emerges from these results is that if the amino acid sequences are 50% identical, or more closely related, it will be possible to build a useful model, by transferring the side chains of the new sequence to the backbone of the most closely-related protein of known structure, retaining the side chain conformation whenever possible. In these circumstances, the model will be expected to have the correct fold in over 90% of the structure, and the overall rms deviation of the backbone will be no more than 1 Å. If the sequences are more closely related, the model will be correspondingly better. Such a model would be of a quality useful for interpreting changes in function.

If the amino acid sequences of the new protein and that of its closest relative of known structure have lower than 50% residue identify, we should be more discouraging about building a useful model. If the sequences have only 20% residue identity, the model might even have the correct fold in only half of the structure, and the atomic deviations of the remaining core might well be over 2 Å. Most people would feel that such a model would not be a useful guide to interpreting the properties of the unknown protein. However, often the binding site of a protein family is better preserved than the rest of the protein structure, and it may be possible to interpret changes in specificity in terms of mutations in and around the binding site itself.

It will have been noticed that our model building procedure has been the most naive and conservative possible: we identify the closest relative of the known structure, and retain as many structural features of this known structure as possible. Many people have suggested that this should be regarded as only a zero-order model, and that more powerful computational techniques might improve it. Such techniques might produce a quantitative improvement in the prediction of the conformation of the common core, or yield useful predictions of portions of the structure outside the core.

To achieve a global improvement in the structure, many people have tried to apply energy minimization or molecular dynamics. These are general methods, based on a detailed quantitative representation of the physical forces involved, to predict the conformations that these forces would create. It has been known for some time that these methods cannot fold up a protein "from scratch". It has more recently become clear that these methods cannot substantially improve a model of the type constructed as we have described. The problem seems to be, that if you take a native protein structure, just as determined by x-ray crystallography, and subject it to energy minimization, the program will find that the experimental structure is not at an energy minimum, and the minimum-energy conformation found will have an rms deviation from the correct structure of about 1 Å. But this is as large or larger than the deviation of the naive model, so significant progress has not been made.

Energy minimization is useful for "tidying up" a structure, for example, closing up gaps in the chain resulting from deletions in the amino acid sequence. But it does not give an effective way to move a model towards the correct structure.

The question of modeling portions of the structure outside the core is one in which some progress has been made, at least for relatively short loops. We have faced the problem of modeling the antigen-binding loops of antibodies [3,4]. Here and in related cases, the effective approach has been to look in the general corpus of known structures for a prefabricated piece that will fit. For hairpin loops between consecutive strands of a β-sheet, there are certain rules relating the conformation of the loops to the length and sequence of the loop [4,5]. In favourable cases, these rules guide us in building the conformation of a loop by stitching in a piece from a known structure.

A more general approach, developed by T. Alwyn Jones and colleagues, is based on a general method of substructure search. This may be thought of as roughly analogous to a standard editing operation: that of identifying occurrences of a character string in text. The basic idea is that if one has fixed two points in the chain, perhaps as the ends of regions of secondary structure in the core, one can extract from a database of known structure examples of regions that match the structure of the two ends, and then can look at what appears in between. In favourable cases, it will emerge that there is a preferred way to connect the two regions and this can be applied to the modeling of the loops.

Conclusions

Advances in experimental techniques have presented us with much knowledge from our biological heritage, in a form accessible to computer data banks and information retrieval systems. The problem we currently face is to provide the interface between the archived data and the practicing scientist, so that this knowledge can be fruitful and multiply.

References

[1] Lesk, A. M., and Chothia, C., The Response of Protein Structures to Amino Acid Sequence Changes, Phil. Trans. Roy. Soc. London **317**, 345 (1986).
[2] Chothia, C., and Lesk, A. M., Relationship Between the Divergence of Sequence and Structure in Proteins, EMBO J. **5**, 823 (1986).
[3] Chothia, C., Lesk, A. M., Levitt, M., Amit, A. G., Mariuzza, R. A., Phillips, S. E. V., and Poljak, R., The Predicted Structure of Immunoglobulin D1.3 and Its Comparison with the Crystal Structure, Science **233**, 755 (1986).
[4] Chothia, C., and Lesk, A. M., Canonical Structures for the Hypervariable Regions of Immunoglobulins, J. Mol. Biol. **196**, 901 (1987).
[5] Sibanda, B. L., and Thornton, J. M., β-hairpin Families in Globular Proteins, Nature **316**, 170 (1985).
[6] Jones, T. A., and Stirup, S., Using Known Substructures in Protein Model Building and Crystallography, EMBO J. **5**, 819 (1986).
[7] Baker, E. N., Structure of Actinidin, After Refinement at 1.7 Angstroms Resolution, J. Mol. Biol. **141**, 441 (1980).
[8] Kamphuis, I. G., Drenth, J., and Baker, E. N., Thiol Proteases. Comparative Studies Based on the High-resolution Structures of Papain and Actinidin, and on Amino Acid Sequence Information for Cathepsins B and H, and Stem Bromelain, J. Mol. Biol. **182**, 317 (1985).
[9] Phillips, S. E. V., Structure and Refinement of Oxymyoglobin at 1.6 Angstroms Resolution, J. Mol. Biol. **142**, 531 (1980).