DNA Systematics

Volume I: Evolution

Editor
S. K. Dutta, Ph.D.

DNA Systematics

Volume I: Evolution

Editor
S. K. Dutta, Ph.D.
Professor
Departments of Botany,
Genetics and Human Genetics,
and Oncology
Howard University
Washington, D.C.

CRC Press, Inc.
Boca Raton, Florida

Library of Congress Cataloging-in-Publication Data
Main entry under title:

DNA systematics.

Includes bibliographies and indexes.
Contents: v. 1. Evolution—v. 2. Plants.
1. Deoxyribonucleic acid—Collected works.
2. Recombinant DNA—Collected works. 3. Chemotaxonomy
—Collected works. I. Dutta, S. K. (Sisir K.)
QP624.D19 1986 574.87'3282 85-21285
ISBN 0-8493-5820-5 (v. 1)
ISBN 0-8493-5821-3 (v. 2)

Direct all inquiries to CRC Press, Inc., 2000 Corporate Blvd., N.W., Boca Raton, Florida, 33431.

International Standard Book Number 0-8493-5820-5 (Volume I)
International Standard Book Number 0-8493-5821-3 (Volume II)

Library of Congress Card Number 85-21285
Printed in the United States

FOREWORD

In the early 1960s in the now defunct Biophysics Section of the Carnegie Institution of Washington, Department of Terrestrial Magnetism, Ellis Bolton, Brian McCarthy and I were amazed that we could readily compare the reassociation of mammalian DNA-RNA and DNA-DNA; this, especially, since current dicta said that, because of great complexity, meaningful reassociation did not take place. Our examinations of unfractionated, sheared eukaryotic nucleic acids were mostly made possible by the presence of repeated families of sequences. These families were soon recognized and analyzed by Roy Britten, David Kohne and Michael Waring. Ultimately, Britten and Eric Davidson produced a theory in which the repeated DNA sequences were used as part of control mechanisms governing transcription functions; testing of this theory is still ongoing. At the same period of time, and before, another Carnegie Institution of Washington Staff Member, Barbara McClintock, was performing genetic experiments with maize which indicated "jumping genes". These experiments, finally recognized by a Nobel Prize, have stimulated a resurgence of interest in DNA control elements which are treated herein along with other forms of transcription translation controls such as the tRNAs and rRNAs. In this book series, an international mix of authors has been assembled to address current progress in control and genome comparison. The next 20 years should provide a very great increase in knowledge of these systems. What will the next related volume in this series contain?

Bill Hoyer

INTRODUCTION

In recent years, numerous studies have been performed that use various characteristics of DNA to estimate the diversity or relatedness between both closely and distantly related species. Some of these studies have been concerned with experimental microevolution dealing with the accumulation of relatively small changes within a species, and some with macroevolution involving taxonomic categories and measurements taken over long periods of time or on a geological scale. Most of this explosion of knowledge has been due to the utilization of the powerful methods of recombinant-DNA technology, and a new interdisciplinary science has evolved spontaneously which may be called "DNA Systematics". Included in this science are the characterization of DNA in nuclear and cytoplasmic genomes; DNA:DNA reassociation kinetics of repeated and nonrepeated DNA sequences; thermal stability measurements of heteroduplexes; restriction enzyme patterns analysis of specific nuclear and nonnuclear DNA segments; rDNAs, and mitochondrial and chloroplast DNAs; the rate of evolution of cell organelle genomes vs. nuclear genomes; the implication of gene duplication and gene fusion in evolution and the evolutionary history of specific genes like rRNA genes and hemoglobin genes; evolutionary trends in regulation; and species specificity and DNA sequencing of processing sites of introns and different RNA maturation sites.

Historically, DNA systematics studies were initiated more than 20 years ago by Ellis Bolton, Roy Britten, Bill Hoyer, David Kohne, Brian McCarthy, and others at the Carnegie Institution of Washington, Washington, D.C., and a few other scientists of England, France, and of U.S.S.R. whose work has been reviewed by Belozersky and Antonov during 1972 and 1980 at Moscow University Press, U.S.S.R. Their techniques were mostly based on DNA:DNA hybridization, which is now claimed as the "most favorable of all" methods for revealing family trees, as discussed by Lewin.* Based on these techniques Charles Sibley and Jon Ahlquist (see footnote) have proposed a "DNA clock" to construct phylogenetic trees. Antonov and his associates from the U.S.S.R proposed similar DNA clocks earlier. These molecular clock(s) are becoming very popular. So much so, that they are "now in danger of becoming the dogma that the fossils once were".* Unfortunately, there is no available comprehensive treatise of this vast amount of new knowledge particularly on microevolution based on studies in DNA systematics using new tools of recombinant-DNA technology. In the present work we attempt the first comprehensive review of new information on DNA systematics.

The enormous amount of accumulated information has been reviewed by authors who are active in their respective areas and then organized into three volumes. Volume I is devoted to general topics of DNA systematics with respect to general evolution. Hobish reviews the present state-of-the-art use of computers for storage and retrieval of DNA research data. The role of movable elements in evolution and species formation is reviewed by Georgiev and his associates. It is well established now that mobile DNA sequences provide variability for natural selection and for evolutionary jumps. It makes genomes flexible, and mobile sequences are widespread in living creatures. Studies have been performed on the evolutionary significance of various control mechanisms which regulate speciation and evolution. Studies dealing with the regulation of ribosomal RNA processing sites, regulation of transcription, and analysis of various small RNAs along with their phylogenetic significance are reviewed by Crouch and Bachellerie; Huang; and Beljanski and Le Goff, respectively. The examination of mitochondrial genomes from mammals, *Drosophila*, and fungi has produced models of mt-DNA variation and offers a comparative treatment of evolution with nuclear genomes; these investigations are reviewed by Birley and Croft. Two of the most important gene sets were selected for discussion in this volume. One is the histone gene set, most genes of which do not have introns, and the other is the hemoglobin gene set,

* Lewin, R., DNA reveals surprises in human family tree: the application of DNA-DNA hybridization, *Science*, 226, 1179, 1984.

which does have introns. Enormous amounts of information on these gene sets on the evolution of *Xenopus*, avians, rodents, and higher primates including humans are reviewed by Marzluff; and Winter, respectively.

The second volume is devoted primarily to the DNA systematics of plants, although, where necessary, reference species other than plants have been included to present a complete story. This volume starts with the classical approach to plant systematics using knowledge obtained from the contents of plant nuclear DNAs; this material is reviewed by Ohri and Khoshoo. Plant species, particularly higher green plants, show polyploidy and have 70 to 75% repeated DNA sequences. Studies made on these repeated and single copy DNA sequences of monocot and dicot plants are reviewed by Mitra and Bhatia; and Antonov, respectively. Appels and Honeycutt; and Troitsky and Bobrova have reviewed extensive studies done on ribosomal RNA genes of plants along with information obtained from other species and have given extensive analysis of phylogenetic significance of these studies. The DNA systematics of some fungal species are reviewed by Ojha and Dutta. The chloroplast genomes of green plants have provided excellent information on plant systematics; this information is reviewed by Palmer, whose group has done extensive work with chloroplast genomes of various plants. A critical glossary of different terminologies used in plant DNA systematics is given by A. K. Sharma.

The third volume, now in preparation in collaboration with Dr. William P. Winter as Co-Editor, is devoted primarily to the DNA systematics and evolution of *Homo sapiens* and related higher primate species. This volume will treat some of the newer insights into relationships within the higher primates and the origin of modern man from anthropoid ancestors. Also included will be discussions of the relationships between the races of man as determined by DNA analysis. In addition, several genes which are of vital concern to humans like neuron-specific genes, lipoprotein genes, HLA genes and others will be discussed. Dr. Ronald L. Nagel of Albert Einstein College of Medicine, New York; Dr. C. G. Sibley of Yale University; Drs. M. G. George, R. M. Millis, Mukesh Verma and S. K. Dutta, and William P. Winter of Howard University; Drs. T. B. Rajaveshisth, A. J. Lusis and others of the University of California at Los Angeles; Dr. I. A. Levedevan of Vavilov Institute of Human Genetics, Moscow, U.S.S.R.; Dr. R. L. Honeycutt of Harvard University; and Dr. R. D. Schmickel of the University of Pennsylvania will be writing chapters for this third volume.

These three volumes are expected to be valuable references, not only to students of evolution but also to others interested in efficient germ plasm resource maintenance and utilization, and fields which are vital for planning plant and animal breeding programs. Knowledge of DNA markers correlating the geographic distribution of genes responsible for heritable diseases such as human sickle cell anemia should be of profound importance to physicians and epidemiologists.

In addition to contributing authors, who have also helped in reviewing several chapters, several other authors have helped in organizing and improving various chapters. I would like to acknowledge particularly Fransciso Ayala of the University of California, Davis; Igor Dawid, H. Westphal and A. Schecter of the National Institutes of Health, Bethesda, Md; Professor A. K. Sharma of Calcutta University, Calcutta, India; R. L. Peterson, George Mathew and D. R. Maglott of Howard University, Washington, D.C.; H. James Price, Texas A&M University, College Station, Texas; H. R. Chen, National Biomedical Research Foundation, Georgetown University Medical Center, Washington, D.C.; Bill Hoyer of Georgetown University; E. S. Weinberg of the University of Pennsylvania, Philadelphia; and G. N. Wilson, Pediatric Genetics, The University of Michigan, Ann Arbor.

S. K. Dutta
Editor

THE EDITOR

Sisir K. Dutta, Ph.D., is Professor of Molecular Genetics in the Department of Botany; and Adjunct Professor in the Departments of Genetics and Human Genetics; and in the Department of Oncology at Howard University, Washington, D.C.

Dr. Dutta obtained his B.S. degree from Dacca University, Bangladesh in 1949, and thereafter, served for 6 years as Research Assistant in Genetics and Plant Breeding for the government of West Bengal, India. He received his M.S. and Ph.D. degrees in genetics from Kansas State University, Manhattan, Kansas in 1958 and 1960, respectively. He was a research associate and/or visiting scientist at the University of Chicago, Columbia and Rockefeller Universities in New York, the Pasteur Institute in Paris, the National Institutes of Health in Bethesda, Maryland, and Rice University in Houston, Texas. He was Chief Research Officer-*cum*-Director of the National Pineapple Research Institute of Malaysia from 1960 to 1964, Chairman of the Division of Natural Sciences, and Chairman of the Biology Department of the Christian University Affiliated College at Hawkins, at Texas Southern University from 1964 to 1967. In 1967 he assumed his present duties at Howard University.

He has been the organizer, chairman, and speaker of several national and international symposia held in the U.S., U.S.S.R., Europe, and Asia. He has been a member of the editorial board of the *East Pakistan Agricultural Journal*, a reviewer and panelist of several government and private agencies. He has been inducted as a personality in America's Hall of Fame for his contribution in molecular genetics, has appeared in *Who's Who in the World*, *Who's Who in America*, and *Who's Who in Frontier Sciences and Technology*. He is a member of several national and international professional societies, author or coauthor of more than 100 papers including monographs, and book chapters and editor of four books. He has been a recipient of several research awards for the U.S. National Science Foundation, National Institutes of Health, Department of Energy, Environmental Protection Agency, Research Corporation, Anna Fuller Fund, and several other agencies including the United Nations Development Projects.

His current research interest is in the areas of regulation of ribosomal RNA transcription and processing, molecular genetics of neuron-specific genes, and molecular evolution.

CONTRIBUTORS

Volume I

Jean-Pierre Bachellerie
Charge de Recherche
Centre Recherche Biochemie and Genetique Cellulaire
Centre National de la Recherche Scientifique
Toulouse, France

Mirko Beljanski
Master in Scientific Research
Department of Pharmacodynamie
Faculté des Sciences Biologiques et Pharmaceutiques
Chatenay-Malabry, France

A. J. Birley
Lecturer
Department of Genetics
University of Birmingham
Birmingham, England

J. H. Croft
Lecturer
Department of Genetics
University of Birmingham
Birmingham, England

Robert J. Crouch
Research Chemist
Laboratory of Molecular Genetics
National Institutes of Health
Bethesda, Maryland

Georgii P. Georgiev
Professor, Head
Department of Nucleic Acid Biosynthesis
Institute of Molecular Biology
U.S.S.R. Academy of Sciences
Moscow, U.S.S.R.

Tatiana I. Gerasimova
Chief
Group of Mobile Elements
Institute of General Genetics
U.S.S.R. Academy Sciences
Moscow, U.S.S.R.

Mitchell K. Hobish
Assistant Research Scientist
Laboratory of Chemical Evolution
Department of Chemistry
University of Maryland
College Park, Maryland

Pien Chien Huang
Professor of Biochemistry
Johns Hopkins University School of Hygiene and Public Health
Baltimore, Maryland

Yurii V. Ilyin
Doctor of Biological Science
Head, Department of Genome Mobility
Institute of Molecular Biology
U.S.S.R. Academy of Sciences
Moscow, U.S.S.R.

Liliane Le Goff
Docteur de Sciences
Department of Pharmacodynamie
Faculté des Sciences Biologiques et Pharmaceutique
Chatenay-Malabry, France

William F. Marzluff
Professor of Chemistry
Florida State University
Tallahassee, Florida

Alexei P. Ryskov
Senior Researcher
Department of Nucleic Acids Biosynthesis
U.S.S.R. Academy of Science
Moscow, U.S.S.R.

William P. Winter
Senior Biochemist
Center for Sickle Cell Disease
Associate Professor of Genetics and Human Genetics and Medicine
Howard University
Washington, D.C.

Volume II

Andrew S. Antonov
Professor
A. N. Belozersky Laboratory of
Molecular Biology
Department of Evolutionary Biochemistry
Moscow State University
Moscow, U.S.S.R.

R. Appels
Principal Research Scientist
Division of Plant Industry
CSIRO
Canberra, ACT, Australia

Chittranjan R. Bhatia
Head
Nuclear Agriculture Division
Bhabha Atomic Research Centre
Trombay, Bombay, India

V. K. Bobrova
Department of Evolutionary Biochemistry
A. N. Belozersky Laboratory of
Molecular Biology and Bioorganic
Chemistry
Moscow State University
Moscow, U.S.S.R.

Sisir K. Dutta
Professor in Molecular Genetics
Department of Botany
Howard University
Washington, D.C.

Rodney L. Honeycutt
Assistant Professor of Biology
Department of Organismic and
Evolutionary Biology
Harvard University
Cambridge, Massachusetts

T. N. Khoshoo
Distinguished Scientist
CSIRO
New Delhi, India

Ranjit K. Mitra
Scientific Officer
Nuclear Agriculture Division
Bhabha Atomic Research Centre
Trombay, Bombay, India

Deepak Ohri
Scientist
Cytogenetics Laboratory
National Botanical Research Institute
Lucknow, India

Mukti Ojha
Maitre d'Enseignement et de Recherche
Biologie Vegetale
Universite de Geneve
Geneva, Switzerland

Jeffrey D. Palmer
Arthur F. Thurnau Assistant Professor of
Molecular Genetics
Division of Biological Sciences
University of Michigan
Ann Arbor, Michigan

Arun Kumar Sharma
Indian National Academy of Science
Professor and Programme Coordinator
Centre of Advanced Study
Department of Botany
University of Calcutta
Calcutta, India

A. V. Troitsky
Department of Evolutionary Biochemistry
A. N. Belozersky Laboratory of
Molecular Biology and Bioorganic
Chemistry
Moscow State University
Moscow, U.S.S.R.

TABLE OF CONTENTS

Volume I

Volume II

Chapter 1

THE ROLE OF THE COMPUTER IN ESTIMATES OF DNA NUCLEOTIDE SEQUENCE DIVERGENCE

Mitchell K. Hobish

TABLE OF CONTENTS

I. INTRODUCTION

The computer has taken on several roles with respect to the analysis of protein and nucleic acid sequences. Indeed, without the computer, we may never have seen the current boom in sequence information. The roles played by the computer cover not only the initial acquisition and subsequent storage and retrieval of sequence data, but also the analysis of those data to determine such features as the location of control regions, structural homology, and estimates of divergence of sequences for the inference of phylogenetic relationships. In this chapter, descriptions of the state-of-the-art in these various areas will be presented, along with some comments and suggestions for optimizing the interaction between the experimenter and the information, as facilitated by the medium of the computer.

Modern scientific instrumentation has revolutionized not only the way in which research is carried out, but even the very research being undertaken. Phenomena whose existence had heretofore only been theorized may now be analyzed with instrumentation capable of measurements over ranges and sensitivities that have improved by several orders of magnitude in just a few years. Concomitantly, there have come significant strides in the ability of computers to capture, store, process, and analyze the data obtained with such instrumentation. Recent advances in the design and manufacture of microprocessors have led to the ubiquity of microcomputers, which have capacities and capabilities that were available only on minicomputers and mainframes only a few years ago. This form of inexpensive, distributed processing augers well for experimental design and execution. As a result, in some cases even an unskilled novice may do in minutes what had previously taken a highly skilled technician several days.[1] The acquisition of data, its analysis, and even its final preparation for presentation may now be almost entirely automated. Analysis of data may be done almost on a real-time basis, with concomitant re-design of experimental protocols as necessary. One result of this process is that the interaction between the investigator and the research underway has been facilitated.

In the specific case of the application of computers and automation to research on nucleic acids, most of the impact has been felt at the data manipulation end of the scale, with comparatively little effect on the acquisition of those data. In the following pages, the role of the computer in such acquisition, as well as the storage and retrieval of nucleic acid sequence data will be examined, with special emphasis on how use of these machines may be used to analyze those data to infer sequence divergence, and hence, phylogeny. This chapter is not designed to present everything you always wanted to know about the inference of phylogenetic trees, but rather a general description of how the computer has been used in that area of research. This topic(s) is presented in other chapters in this volume.

II. COMPUTERS IN THE ACQUISITION OF NUCLEOTIDE SEQUENCE DATA

The development and refinement of rapid techniques for DNA sequencing[2,3] have contributed directly to the present boom in nucleic acid sequence research. Also contributory has been the availability of low-cost computational facilities to the laboratories where the sequencing is being done. This local capability has given rise to systems that aid in the acquisition of sequence data and its subsequent preliminary analysis.

As a data acquisition tool, the computer acts as an extension of the bench techniques which generate raw sequence data as autoradiographs. Several programs have been written[4-7] which enable the operator to directly read DNA sequences from such gel autoradiographs using a graphics tablet (a digitizer) as a transducer. In this way the operational information is converted to a form directly accessible by a computer, with subsequent storage for later manipulation. Errors in reading the gels or in typographical transcription of the sequences are obviated. Such programs offer the operator several options beyond data

transduction, including, but not limited to, portrayal of the sequence and verification thereof (by replicate scanning of the gel), and the ability to write the sequence data to a mass storage device for later manipulation or transmission to another program or computer. In addition, by carefully defining the way in which the gels are read, irregularities in spacing of the bands are included in the data set. These irregularities arise due to compression of the gel as a result of the interaction between regions of secondary structure induced in the DNA fragments and the gel matrix.[5] This information itself may be of use in the investigation of such structures.[8,9] A necessary feature in accurate transcription of gel data is real-time quality control on the part of the operator. The program should provide the operator some means of ensuring that the base entered into the computer is correct from the standpoint of reading the gel. For example, Lautenberger[4] uses a simple audio feedback system, while Gingeras et al.[5] employ recent advances in the area of computer-generated speech to actually tell the operator what is happening.

Once sequences have been determined, whether aided by computer or not, true manipulation of those data may begin. These operations may range from simple calculation of base composition, to translation of the sequence in several reading frames, and pattern matching in an attempt to identify regions of interest such as palindromes,[10] control regions,[11] and restriction sites,[12-14] in which information may be used in further restriction analysis.[15] Such utility is subject to several constraints, as described by Gallant,[16] who demonstrates how sequence reconstruction by proper alignment and overlap of cleavage products is "computationally intractable" for large numbers of input data. The problem may be handled, however, by defining several boundary conditions on the input data set, which leads to a limited number of candidate sequences. This limited set may now be used for further sequence refinement using cleaving agents with increasing specificity. Except for this proviso, these programs may be used on a real-time basis to determine what sequencing strategy may next be followed. This real-time basis defines an apparent need for local computing power; desktop or microcomputers are ideally suited for this purpose, except for some constraints on the length of sequences with which a user deals. In contrast to the data acquisition level of use, the type of computer used (or, more exactly, the clock speed of the computer), the amount of memory available, and the implementation of a specific algorithm will provide the limiting factors. For this reason, only relatively short sequences may be dealt with, unless the programs are implemented on a mini- or mainframe system. Since the original sequence data are obtained in relatively short pieces, this is, at least at this stage, no hardship. Similarly, from the standpoint of rapid pattern recognition, interpreted BASIC is not well suited here. Since the sequences may often be treated as strings of characters, a language without adequate string handling capacity would also limit efficiency. On the other hand, a language built around operations involving strings, arrays, or tables could reduce processing time. A language well suited for such array analysis is APL. Unfortunately, this language is not often implemented on microcomputers. Furthermore, the language suffers from a case of terminal unreadability with increasing temporal distance from the creation of a program.

Indeed, it is the string-of-characters representation of sequence data which provides a complicating factor in the presentation of the raw data to the investigator. Representing the sequence data on a CRT or hard-copy device as a string of bases or base symbols may limit the utility of the data, since there is no obvious delineation of regions of interest. There has been at least one attempt to overcome several limitations inherent in string representation of nucleic acid sequence data,[17] but this seems to have its greatest utility in examining the overall structure of large sequences. Similar approaches to this problem have been taken by Staden,[18] who presents the output of a program (ANALYSEQ) showing alignments between protein coding regions, splice junctions, ribosome-binding sites, and poly-A sites.

The increasing appearance of inexpensive medium and high resolution color graphics terminals provides an interesting possibility for representation of pattern analysis of nucleic

acid sequences. For example, regions of interest, as mentioned above, could all be delineated in different colors. This technique has been taken virtually to state-of-the-art for microcomputers with a program described by Watanabe et al.[19] In this paper, the authors describe a microcomputer program which produces a three-dimensional, color display of protein and nucleic acid structure. Rotation and enlargement of defined regions may also be done. Eventually it may be possible to compare nucleic acid sequence structures by superposition and cancellation of homologous regions, leaving only the difference structures for comparison.

The graphic representation of sequences and/or structures notwithstanding, several programs which aid in pattern recognition have been written. The more complicated of these[20-23] have such extensive sequence handling options that they must be implemented on a mainframe system. For example, the packages mentioned above have been used on DEC System 10 and 20 systems, and are being licensed to the commercial community by Intelligenetics. These programs, and others like them,[24,25] are generally run in an interactive mode and allow rapid pattern analysis of sequences whose lengths prohibit analysis on memory-limited microcomputers. The search for rapid and efficient analysis by character pattern recognition continues, and has recently been addressed by several groups.[26-28] However, it should be noted that while pattern recognition is possible by computer analysis, it is very often the human aspect of the system which can recognize patterns that may be just barely above noise. Recent advances in artificial intelligence and its application to "expert systems" will certainly have an impact in this area.

III. NUCLEIC ACID SEQUENCE DATABASES

Historically, even as the amount of sequence data was growing, the research community began to recognize a need for a centralized, standardized sequence data base. Despite reports of sequences as early as 1978, the first of these, the Nucleotide Sequence Data Library of the European Molecular Biology Laboratory was not established until April 1982. A similar facility for U.S. researchers had been considered as early as 1979, and by the end of that year Dayhoff's group (NBRF, the National Biomedical Research Foundation)[29] in Washington, D.C., and Goad's at Los Alamos,[25,30] had submitted proposals for development to the National Institutes of Health (NIH). Many issues had to be dealt with, including security of "sequences in progress", and the need to determine the most effective means of implementing the data network. Even the Department of Defense became involved, since the ARPANET was to be used as the common information nexus. An early attempt to centralize research resulted in the GENET package.[31] The package was implemented on the SUMEX-AIM (Standford University Medical Experimental computer-Artificial Intelligence in Medicine) system and demonstrated the utility of artificial intelligence methods to a "knowledge-based" information system. One advantage of this method is the ability to apply sophisticated pattern recognition algorithms to large numbers of sequences, thereby increasing the utility of the computer in nucleic acid research by test cloning (simulating) sequencing strategies in advance of an actual experiment.[32] Unfortunately, the desire for access to the information available on SUMEX through GENET eventually resulted in complete shutdown of that account. Individual users were left to their own devices, with the result that there has been significant redundancy of effort in the generation of sequence analysis software.

Eventually, Bolt Beranek and Newman, Inc. (BBN), a private corporation located in Cambridge, Mass., with expertise in computer communications, won a contract to develop a national sequence database,[33] now called GenBank™, the Genetic Sequence Data Bank. GenBank, a trademark of the NIH, is a U.S. government-sponsored internationally available repository of all reported nucleic acid sequences greater than 50 nucleotides in length, cataloged, and annotated for sites of biological interest and checked for accuracy. GenBank™ was created by the National Institue of General Medical Sciences of NIH in 1982. Co-

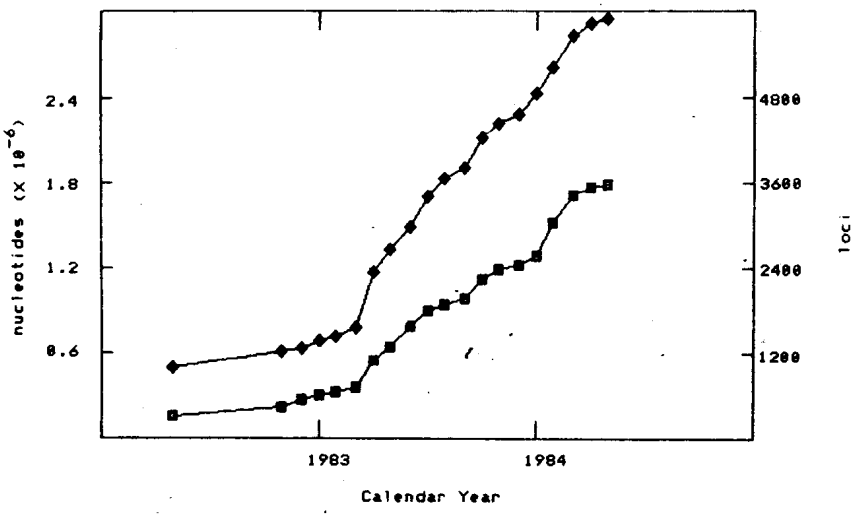

FIGURE 1. Growth of GenBank™, 1982 to the present. Data were obtained from Bolt Beranek and Newman, Inc. Diamonds (◆) represent the number of nucleotides; squares (■) represent the number of loci.

sponsors include the National Cancer Institute, the National Institute of Allergy and Infectious Disease, the Division of Research Resources of NIH, the National Institute of Arthritis, Diabetes and Digestive and Kidney Diseases, as well as the National Science Foundation, the Department of Energy, and the Department of Defense.

Despite this investment, recently the NIH awarded $5.6 million (1985) to IntelliGenetics of Palo Alto, Calif., for the purpose of establishing a national computer resource for molecular biology.[34] The database, to be called Bionet, will provide a centralized resource of nucleic acid and protein sequences, and provide sophisticated software for manipulation of the data. In addition, it is hoped that the network would provide an electronic community of investigators, which could facilitate communication between research groups about topics of common interest. The IntelliGenetics effort is an outgrowth of the GENET network, but conceived of and executed with regard for the problems encountered by its predecessor. For example, the number of researchers accessing Bionet will start at 10, and may be increased to 15, as compared with GENETs 2 lines. Furthermore, Bionet will be accessible via local numbers from all states. The users will comprise about 300 research groups and perhaps 1000 individual subscribers, all linked in a loose confederation of investigators in a huge "laboratory without walls".[35] Thus, in addition to providing the sequence information itself, the network will provide several of the functions of electronic conference call, allowing routine and essentially instantaneous communication between researchers across the country.

Despite the lack of a unique centralized resource there has been tremendous growth in the number of sequences which have been reported and stored in the several databanks currently available. For example, at its inception in 1978, the NBRF database had some 11 sequences, comprised of 24,658 nucleotides. By January 1984, this number had risen to 1234 sequences, of 1,822,759 nucleotides. Impressive as this is, it is eclipsed by the data for GenBank: as of April 1984, there were 3576 loci recorded, consisting of 2,945,001 nucleotides. The growth rate is evident from Figure 1.

As an example of the type of information available in these databanks, we shall examine a few features of GenBank. The information in GenBank is provided in the form of 11 tables, of 31 columns, and as many rows as there are entries in the database. The 31 columns, contain such information as the locus (named), a short description of the locus, the actual sequence, the published reference, annotated structural regions, and more. A typical entry,

```
GETGENBANKENTRY<GO>
[ACCESSING GENBANK ENTRIES ...DONE]

WHICH LOCUS DO YOU WISH TO SEE?RABHBA<GO>
[RETRIEVING  RABHBA ...DONE]

Annotated listing for sequence RABHBA  3/11/84

DEFINITION:
RABBIT ALPHA-GLOBIN MRNA.

              10         20         30         40         50
               :          :          :          :          :
    1 ACACTTCTGG TCCAGTCCGA CTGAGAAGGA ACCACCATGG TGCTGTCTCC
   51 CGCTCACAAG ACCAACATCA AGACTGCCTG GGAAAAGATC GGCAGCCACG
  101 GTGGCGAGTA TGGCGCCCAG GCCGTCGAGA GGATGTTCTT GGCCTTCCCC
  151 ACCACCAAGA CCTACTTCCC CCACTTCGAC TTCACCCACG GCTCTGAGCA
  201 GATCAAAGCC CACGGCAAGA AGGTGTCCGA AGCCCTGACC AAGGCCGTGG

             260        270        280        290        300
               :          :          :          :          :
  251 GCCACCTGGA CGACCTGCCC GGCGCCCTGT CTACTCTCAG CGACCTGCAC
  301 GCGCACAAGC TGCGGGTGGA CCCGGTGAAT TCAAGCTC  TGTCCCACTG
  351 CCTGCTGGTG ACCCTGGCCA ACCACCACCC CAGTGAATTC ACCCCTGCGG
  401 TGCATGCCTC CCTGGACAAG TTCCTGGCCA ACGTGAGCAC CGTGCTGACC
  451 TCCAAATATC GTTAAGCTGG AGCCTGGGAG CCGGCCTGCC CTCCGCCCCC

             510        520        530        540        550
               :          :          :          :          :
  501 CCCATCCCCG CAGCCCACCC CTGGTCTTTG AATAAAGTCT GAGTGAGTGG
  551 CA

** Composition**

    113  A
    198  C
    144  G
     97  T
Length = 552

TYPE:
MRNA
```

FIGURE 2. Sample entry in GenBank®. (Courtesy of Bolt Beranek and Newman, Inc.)

in this case, for rabbit hemoglobin α mRNA, is reproduced in Figure 2. The row and column structure enable rapid access and display of information. For example, the command:

DISPLAY PHAGESEQUENCES ROWS 1 TO 10 COLUMNS 1 TO 5

would result in the information found in Figure 3.

The entire database is found in the table GENBANK. The remaining ten tables contain grouped sequences: mammalian sequences, other vertebrate sequences, invertebrate sequences, plant sequences, organelle sequences, bacterial sequences, structural RNA sequences, viral sequences, bacteriophage sequences, and synthetic sequences. As of April 1984, the 3576 loci presented arise from 4471 individual reports.

Recently, BBN has provided an alternative means of manipulating the databank through the medium of "user-friendly" menus. The user is guided step-by-step through the process, and helpful hints and reminders about response format are provided throughout the menu system. At all stages, help about the menu under scrutiny is available, so there appear to be few places where even a neophyte could get into serious trouble. This could be of great utility when it comes to minimizing on-line charges, which, as of April 1984, amounted to