Report 83-51
Stanford -- KSL

The Role of Computers in Biotechnology.
Peter E. Friedland, Paul Armstrong,
Thomas Kehler,
Sep 1983

card 1 of 1

Heuristic Programming Project
Report No. HPP 83-51

September 1983

The Role of Computers in Biotechnology

Peter E. Friedland, Paul Armstrong, and Thomas Kehler

Published in the Bio/Technology, September 1983.

The Role of Computers in Biotechnology

Peter E. Friedland
Computer Science Department
Stanford University
Stanford, California 94305

Paul Armstrong and Thomas Kehler
IntelliGenetics, Inc.
124 University Avenue
Palo Alto, California 94301

1. Introduction

The revolutionary advances of the early 1970's, rapid nucleic acid sequencing, restriction enzyme technology, and recombinant DNA techniques, have produced an information explosion in the field of molecular biology. Millions of base pairs worth of DNA sequences are now known and must be analyzed; hundreds of restriction enzymes and cloning vectors must be kept track of to carry out work in genetic engineering. Computational systems have become a necessary tool to acquire, retrieve, manipulate, and analyze this information.

A myth of the early days of computing was that computers were excellent number manipulators, but poor at working with other forms of information. The truth, which has emerged from two decades of work in the field of artificial intelligence, is that computers are excellent, general-purpose *symbol* manipulators. Numbers are a type of symbol, but so, too, are nucleic acid sequences, restriction maps, and even more complex forms of knowledge like strategies for carrying out recombinant DNA experiments.

This article will survey current applications of computers to biotechnology. It will also discuss methods of delivery of computational tools to working scientists. Finally, it will take a look at the future of computing in the biotechnology field with an emphasis on applications to problems that occur during the scale-up and product testing and delivery phases of industrial biotechnology.

2. Current Applications to Biotechnology Research

Current applications can be broken up into the following broad classes of computational tools: data collection, assembly, storage and retrieval, nucleic acid and protein sequence analysis, restriction map generation, simulation of molecular processes, and experiment planning and debugging. Each of these classes will be described below. (An excellent reference source for current applications is the special issue of *Nucleic Acids Research* of January 11, 1982 devoted exclusively to the subject of computers in molecular biology.)

2.1. Data Collection, Assembly, Storage, and Retrieval

An important requirement for all scientific research is the ability to collect and store the primary data of the particular field; in the case of biotechnology, the primary data consists of nucleic acid and protein sequences. Computer programs are now in widespread use to facilitate all aspects of sequence data collection, verification, and correction. These programs include a variety of editing systems that have the power to take advantage of several "smart" display terminals. In addition, several specialized terminals for sequence entry have been developed; one example is the IntelliGenetics Sequen^{TM} shown below.

An interesting computational problem occurs as a result of the current limitations on DNA sequencing; a maximum of 400 base pairs per sequencing gel is normal. The task is to combine the results of many sequencing gels into a single long sequence. It involves finding overlaps, choosing the best overlaps, allowing scientists to verify or correct their primary data, and combining the overlap into a single "consensus" sequence of the molecule being studied. Several computer programs have been developed to automate this extremely time-consuming process.

Currently, there are several major efforts underway worldwide to provide archival-quality databases for both nucleic acid and protein sequences. The best known of these are the NIH-supported GENBANK (run by BBN, Inc. and Los Alamos Scientific Laboratory and containing well over a million base pairs of sequence data), the European Molecular Biology Laboratory nucleic acid sequence database, and the National Biomedical Research Foundation nucleic acid and protein sequence databases.

As the databases grow, the information retrieval problem becomes greater. Computer program now exist to search the databases in various ways. The search can be on a textual level; for example, to locate for further analysis all hormone genes across several organisms. The search can also be for various biological patterns; for example, intron-exon junctions, putative promoter sites, or splicing sites in nucleic acids and drug interaction or secondary structure sitesin proteins. The more sophisticated of these programs allow the biologists to describe patterns of interest in close to "biological English."

Storage and retrieval of other biological information is becoming increasingly important. This information ranges from other factual information like restriction enzyme cleavage sites and cloning vector availability and utility to more procedural and strategic knowledge like *heuristics* (rules-of-thumb) for cloning into Lambda phage. See the final section of this article for a discussion of the artificial intelligence-derived concept of an "intelligent encyclopedia" for general knowledge management.

2.2. Nucleic Acid and Protein Sequence Analysis

Many important analytical functions are now being done routinely by computer on nucleic acid and protein sequences. For nucleic acids, these functions include:

- determination of mono-, di-, and tri-nucleotide frequencies
- location of restriction enzyme cleavage sites (as well as other sites that can be described as a pattern of nucleotides)
- location of pair-rich regions of various types
- location of homologies, symmetries, and dyad symmetries both within and among molecules
- translation to protein sequences
- determination of secondary structure for RNA's

For proteins, these functions include:

- determination of amino acid frequencies
- reverse translation of nucleic acid sequence
- location of possible restriction enzyme cleavage sites in the corresponding nucleic acid sequence
- determination of secondary structure (alpha helices, beta sheets, and turns)

- location of hydrophobic and hydrophilic regions

- comparison (both structural and functional) with other protein sequences

2.3. Restriction Map Generation

Clearly, restriction maps can be unambiguously generated from nucleotide sequences; indeed the computer programs described in the previous section include that generation as one of their most-often-used functions. However, a normal early component of many genetic engineering experiments, prior to the availability of complete sequence data, is the determination of a restriction endonuclease map from single, double, and (sometimes) partial and triple enzyme digest data. This determination is normally combinatoric; in all but the simplest of cases there are many millions of possible maps that must be considered to ensure that the best map has been chosen. A computer program called MAP^{TM} has been developed by collaboration between Dr. William Pearson of Johns Hopkins University and IntelliGenetics. MAP calculates digest lengths directly from band positions on gels, takes into account constraining information like the existence of known sites and fragments, and presents a ranked list of all plausible maps to the experimenter.

2.4. Simulation of Molecular Processes

Computers have been used to simulate many aspects of biological systems. Two examples from the MOLGEN project at Stanford University illustrate the range of simulated processes. The first was on a strictly structural level. A problem in early cloning experiments was to maximize the percentage of single inserts of a linear fragment into an opened plasmid which would then recircularize. A straightforward Monte Carlo simulation system was written, based upon the theoretical work of Dugaiczyk et. al., which determined optimal initial concentrations of reactants and length of reaction time.

The second example is more recent work to simulate the structure-to-function relationship of the Lambda phage system as it "decides" whether to act in a lytic or lysogenic manner. The artificial intelligence-based *knowledge base* systems of MOLGEN were used to build a completely transparent simulation system which has biologist-supplied knowledge about genes, protein products, and control loci in Lambda. The system correctly predicts the behavior of Lambda when various combinations of genes and control loci are mutated (including predictions of uncertainty and instability). The knowledge-based approach allows computer-naive biologists to examine and alter the biological information used by the simulation system without becoming computer experts.

2.5. Experiment Planning and Debugging

The problem of planning and debugging both synthesis and analysis experiments is a difficult cognitive task. Experiment planning entails choosing a sequence of laboratory tasks to satisfy some experimental goal. Experiment debugging occurs in the inevitable case when something does not work properly the first time. The MOLGEN project, a joint effort among computer scientists and molecular biologists at Stanford University, is an eight-year effort to apply artificial intelligence to assist scientists in solving these problems.

One discovery of the MOLGEN work is that scientists rarely plan experiments from scratch. Instead they begin with an abstract or "skeletal" plan that is refined to a particular experiment by choosing specific laboratory objects or techniques to "instantiate" each plan step. For example, all cloning experiments may be thought of as arising from the following skeletal plan:

1. isolate the DNA to be cloned
2. choose a cloning vector
3. insert the DNA into the cloning vector
4. choose a host for the new hybrid molecule
5. insert the hybrid into the host
6. select for the clones of interest

The important point is that a scientist, when designing a new cloning experiment, does not reinvent the basic strategy described above. He uses his knowledge of vectors, hosts, insertion methods, and so on, to pick the most applicable and optimal current methods to refine the plan for his particular cloning problem. A general purpose system, known as SPEX, has been constructed that uses the skeletal plan method for design. SPEX makes use of a biologist-described *knowledge base* to become expert in a particular area of molecular biology. It is currently in use within the cloning experiment design field.

Experiments can fail for one of three basic reasons:

1. A technical error was made--the scientist added too much salt or did not run the centrifuge at the correct speed.

2. A technique selection error was made--the scientist picked an exonuclease when he really wanted an endonuclease or he forgot that chaotropic solvents interact unfavorably with certain enzymes.

3. The overall plan was faulty--the individual component steps of the experiment plan worked, but the end result did not solve the original problem.

An artificial intelligence system has been recently completed in the MOLGEN project to help a scientist to analyze faulty experiments for these three types of problems. It uses SPEX to determine if design choices are valid, and simulates the actions of each step to predict changes in key laboratory parameters that may be compared to actual results to find technical errors and isolate the exact nature and location of the flaw.

3. Delivery Methods for Computational Tools

There is a revolution occurring in modern computing. Hardware is becoming simultaneously smaller, cheaper, and more powerful. As that occurs, delivery methods of computational tools for biotechnology are rapidly changing. There are three basic ways of providing scientists with such tools: timesharing, general-purpose single-user computers, and customized biotechnology workstations.

Throughout the 1970's and the early 1980's, timesharing was the most efficient and economical way to provide compuational resources to large groups of scientists. The Digital Equipment Corporation (DEC) PDP-10^{TM}, System 20^{TM}, and VAX^{TM} series of computers are the workhorse machines for scientific timesharing in molecular biology. For example, IntelliGenetics, Inc. operates a DEC 2060 machine which serves several hundred individual molecular biologist users. Timesharing on a shared, central resource is still an effective way to introduce scientists to the power of computational tools for a very low initial entry cost.

The availability of inexpensive, relatively powerful mini and micro computers has made it possible for many molecular biologists to purchase their own, dedicated, single-user computers. These span an enormous range of cost, computing power, and software availability and usability.

Finally, a recent trend in professional computing is the emergence of the customized workstation computer. This is a machine that, for all purposes, can be treated as another laboratory instrument like a centrifuge or scintillation counter. It comes packaged as a hardware-software hybrid with specially designed interfaces that allow molecular biologists to communicate in their own scientific language. An example of such a workstation is the IntelliGenetics BION^{TM} shown below.

4. The Future of Computers in Biotechnology

Biotechnology and computing are both changing rapidly. Computing hardware will continue to get less expensive, smaller, and more powerful. Software will continue to get more expensive as the problems become harder and the people costs rise. Artificial intelligence systems will start to play increasingly important roles in managing the most precious component of a high technology enterprise, the expertise of its scientists and engineers. Some potential emerging applications are discussed in the final paragraphs of this article.

4.1. Advanced Personal Workstations for Scientists

Approximately ten years ago computer scientists at Xerox's Palo Alto Research Center designed a personal computer system which changed the basic nature of human interaction with computers. The design goal for this system was to provide a highly interactive graphics interface to the user. One might think of the computer in this sense as a dynamic "workbench" which could display information in a wide variety of forms, allowing the user to point, make comments, sketch figures, and move chunks of information about on the screen.

Today, systems which use bitmapped graphics and pointing devices such as a mouse or lightpen are fairly common. However, few application systems have fully exploited the capabilities of these kinds of systems. The ability to design specialized graphics-based languages for scientists is one area that shows significant promise. The BION^{TM} shown previously contains such a biologist-specific language which allows a scientist to communicate effectively with the system without learning arbitrary computer languages.

The future of scientific workstations holds great promise to tighten the link between human and computer. There is an increasing need to more tightly couple human problem-solving methods with software systems. Recent research in artificial intelligence has produced the concept of the *knowledge-based system* to provide such a link. Both factual and procedural knowledge, along with the strategies for reasoning with the knowledge, are represented in a manner natural to the human domain expert. Relatively simple reasoning methods, combined with large amounts of expert knowledge, have succeeded in providing expert or near-expert performance in areas ranging from molecular structure elucidation to antibiotic prescription to computer design and repair.

Significantly, the user of a knowledge-based system is able to ask how a conclusion was reached and receive an explanation which delineates specific facts, judgements, and associations used in to reach a particular conclusion. This allows experts to "debug" knowledge bases interactively and the systems to begin to serve as knowledgeable problem-solving assistants.

New computer architectures have been developed which support knowledge-based systems. Example systems currently available on the market are the XEROX 1100^{TM} series of Scientific Workstations and the Symbolics, Inc. 3600^{TM} Workstation. These systems combine facilities for carrying out artificial intelligence research with sophisticated interactive graphics.

4.2. The Intelligent Encyclopedia

One class of applications for knowledge-based systems is the "intelligent encyclopedia." An intelligent encylopedia allows a knowledge base to become an advisory system or it can provide a facility to "play out" the consequences of a strategy. In the figure below we show an intelligent encyclopedia for the berries of California running on the XEROX 1100 Scientific Workstation. BERRY allows the user to either view information about berries and their habitats or to "activate" the knowledge and have the system become an expert advisor on a particular berry. BERRY was constructed using the IntelliGenetics KEE^{TM} system. KEE, the Knowledge Engineering Environment, provides a software facility for building intelligent encyclopedias or other knowledge-based systems.

Figure 4-1: An Intelligent Encyclopedia for California Berries

Intelligent encyclopedias have also been applied to biotechnology. One example system built at IntelliGenetics provides an experiment simulation environment for cloning experiments. A wide variety of DNA structures and proteins are represented in terms of attributes relevant to experiment planning goals. The representation of these structures can be easily changed by the user of the system. The knowledge base contains sets of rules, described directly by a molecular biologist in close to "biological English," for generating restriction maps and simulating recombinant DNA experiment steps. The user can browse through the knowledge base of structures and strategies and new structures can be generated by carrying out a simulated experiment. The figure below shows a screen image of the knowledge base and a restriction map generated by carrying out such an experiment.

Figure 4-2: An Intelligent Encyclopedia for Cloning Experiments

4.3. Design Aids For Scale-up

The ability of fermentation engineers to apply technical knowledge to problems in process modeling and process scale-up is of fundamental importance to the development of the biotechnology industry. Biochemical processes are immensely complex, involving theoretical and empirical knowledge from microbiology, molecular genetics, and biochemical engineering. The crucial empirical knowledge tends to be judgmental, derived by an expert after years of experience.

Heuristic knowledge, typified by good judgment or good guessing, is an integral part of expert performance in fermentation development and control.

Expert performance in developing fermentation processes can be enhanced by providing a supportive computational environment for the problem-solving tasks undertaken by the human expert. Using advanced scientific workstations, the expert can build knowledge bases which capture elements of the his experience. Knowledge bases of this type go beyond the formal mathematical knowledge used in process modeling. The focus of knowledge representation is to capture the informal qualitative judgements which differentiate the seasoned expert from the novice. Capturing this knowledge and making it available to both the expert himself and less experienced colleagues has the potential of providing significant improvements in the speed at which processes are brought from theoretical design to actual productive use.

As an example application, consider an intelligent encyclopedia on metabolic reactions. The knowledge base consists of metabolite descriptions, enzyme descriptions, process descriptions, feedback control mechanisms, and related objects or processes. The focus of the encyclopedia is not to model each step in the metabolic pathway, but to model the important dependencies of products and intermediates on enzymes and other products. Thus, a process model in the knowledge base would be a collection of descriptions and important data dependencies. The goal for the knowledge base description is to have summarized process models which can be used to simulate various metabolic reactions.

Knowledge-based systems will provide advanced planning and development environments for fermentation designers in much the same way that knowledge-based computer-aided design systems for integrated circuits are being used today. Much of the practical experience-based knowledge that relies on qualitative descriptions of processes will be captured in systems that are designed to reflect the engineer's problem-solving needs.

4.4. The Biotech Plant of the Future

As microprocessors decrease in cost and increase in power, the prospect of highly automated systems for production in biotechnology becomes more likely. Knowledge-based systems, coupled with automated fermentation control systems, will allow the biochemical engineer to directly communicate control strategies for processes to the fermentation plant control system. Not only will these strategies be communicated in terms of language that is natural to the biochemical engineer, but the systems will also be able to explain why particular actions were taken or how they diagnosed a particular fault condition in a process. Future biotechnology pilot plants will provide biochemical engineers with highly improved interfaces to data. By combining this improved access to data with the capability to tune control strategies directly, it is likely that complex processes can be understood and controlled much more quickly than at present.

Table of Contents

List of Figures

Biotechnology

ly that French products will
significant impact on interna-
biotechnical pharmaceutical
s. The departure of Jean-
Chevenement may also hurt
nology. As Minister of Indus-
Research he avidly supported
re ambitious proposals of the
nology mobilization program.
gnation coincided with the
cement of further cuts in gov-
t spending, which may bode
rench biotechnology research.
ough France, West Germany,
e U.K. provide much of the
tum in biotechnology re-
Europe's leading biotechnolo-
pany is based in Denmark. In
Novo, a small enzyme and
aceuticals manufacturer locat-
Copenhagen, decided to seek
capitalization. The relatively
anish capital market forced
pany to seek funding over-
ventually they made substan-
lic offerings and are now list-
he New York Stock Exchange;
also one of the fastest grow-
technology companies in the

d in Switzerland, Biogen is Eu-
leading venture company. Like
t has successfully raised capital
rnational markets; Biogen was
t European venture company
r shares to the general public.
ectives are considerably more
tious than those of most venture
anies. It intends to enter the
maceutical market as a fully inte-
d manufacturer, researching,
oping, and marketing its own
ucts.

far, Biogen's research in bio-
nology outside the pharmaceuti-
field has been disappointing, an
nexpected setback. In 1981 Biogen
redicted that many of the real gains
biotechnology would be not in
pharmaceuticals, but in agriculture
nd chemicals. Although these appli-
cations have been realized more slow
ly than anticipated. Biogen is well
placed to exploit these markets when
they do emerge.

Two other European countries are encouraging rapid development of biotechnology. Stringent regulations in the Netherlands hampered commercialization of biotechnology, but the government is now moving to support its domestic industries. A seven year, 75 million guilder biotechnology innovation program has been established, with most of the funding going to established Dutch companies such as Shell, Akzo, and Gist. Gist is a leader in conventional biotechnology—the company produces 25 percent of the world's crude penicillin and is a major enzyme manufacturer. The company recently announced plans to cut back basic pharmaceutical research in favor of product development. This confirms the presumption, held by much of the Dutch pharmaceutical industry, that only governments or large multinationals can endure the costs of basic drug research.

Biotechnology is also seen as an aid to Spain's flagging drug industry. A major market for many years, the Spanish drug industry has shrunk precipitously, with the closure of some 30 firms. Strong overseas competition, punitive price controls, and overly strict regulatory procedures are all in part to blame. The government has moved to restore some balance to the industry by financing six projects on drug-related biotechnology through the Industrial and Technological Research Council. In collaboration with the private company Laboratorios Sobrino, the government is also funding Ingenesa, one of Spain's two venture companies. The specific research targets for this company have not been announced, but there will be a strong emphasis on rDNA technology. The other Spanish venture company, Alter-Llorente, is entirely industry sponsored, with funding coming from Laboratories Alter and Instituto Llorente. Early plans include the building of a $1 million facility for the manufacture of leukocyte interferon using U.S. technology.

The market for biotechnical pharmaceuticals will more than double in the next decade, but it is not yet clear that European laboratories and drug houses will have developed the advanced biotechnology to produce the new products. Hopefully, many of the problems outlined above will disappear in the next few years. If major economic problems persist, the differences in the national markets may begin to diminish. This will be a relief to many, as Europe is a fast-growing market made infinitely more complex by national fragmentation.

John Gurnsey is senior market research analyst for Creative Strategies International, Banda House, Cambridge Grove, London, W6 0LN U.K. He recently published a study of Western European biotechnology in the pharmaceutical industry.

KNOWLEDGE BASE SYSTEMS

THE ROLE OF COMPUTERS IN BIOTECHNOLOGY

The revolutionary advances of the early 1970's, such as rapid nucleic acid sequencing, restriction enzyme technology, and recombinant DNA techniques, have produced an information explosion in the field of molecular biology. Millions of base pairs of DNA sequences are now known and must be analyzed; hundreds of restriction enzymes and cloning vectors must be kept track of for work in genetic engineering. Computational systems have become necessary tools for acquisition, retrieval, manipulation, and analysis of this information. In the early days of computing, it was believed that computers were excellent number manipulators, but poor at working with other forms of information. The truth, which has emerged from two decades of work in the field of artificial intelligence, is that computers are excellent, general-purpose symbol manipulators. Numbers are a type of symbol, but so, too, are nucleic acid sequences, restriction maps, and even more complex forms of knowledge like strategies for carrying out recombinant DNA experiments.

This article will survey current applications of computers to biotechnology. It will also discuss methods of delivery of computational tools to working scientists. Finally, it will take a look at the future of computing in the biotechnology field with an emphasis on applications to problems that occur during the scale-up, product testing, and delivery phases of industrial biotechnology.

Current Applications to Biotechnology Research

Current applications can be broken up into the following broad classes of computational tools: data collection, assembly, storage and retrieval, nucleic acid and protein sequence analysis, restriction map generation, simulation of molecular processes, and experiment planning and debugging. Each of these classes will be described below. (An excellent reference source for current applications is the special issue of *Nucleic Acids Research* of January 11, 1982 devoted exclusively to the subject of computers in molecular biology.)

Data Collection, Assembly, Storage and Retrieval. An important requirement for all scientific research is the ability to collect and store the primary data of the particular field; in the case of genetic engineering, the primary data often consists of nucleic acid and efficient and economical way to provide computational resources to large groups of scientists. The Digital Equipment Corporation (DEC) PDP-10™, System 20™, and VAX™ series of computers are the workhorse machines for scientific timesharing in molecular biology. For example, IntelliGenetics, Inc. operates a DEC 2060 machine that serves several hundred molecular biologists. Timesharing a central resource is still an effective way to introduce scientists to the power of computational tools for a very low entry cost. More recently the availability of inexpensive, relatively powerful mini- and micro-computers has made it possible for many molecular biologists to purchase dedicated, single-user computers. These span an enormous range of cost, computing power, and software availability and usability.

Another recent trend in professional computing is the emergence of the customized workstation computer. This is a machine that can be treated as another laboratory instrument like a centrifuge or scintillation counter. It comes packaged as a hardware-software hybrid with specially designed interfaces that communicate with molecular biologists in their own scientific language. An example of such a workstation is the IntelliGenetics BION™.

Advanced Personal Workstations for Scientists. Approximately 10 years ago computer scientists at Xerox's Palo Alto Research Center designed a personal computer system that changed the basic nature of human interaction with computers. This system provided a highly interactive graphics interface to the user. One might think of the computer in this sense as a dynamic "workbench" that could display information in a wide variety of forms, allowing the user to point, make comments, sketch figures, and move chunks of information about on the screen. Today, systems that use bitmapped graphics and pointing devices, such as a mouse or lightpen, are fairly common; however, few application systems have fully exploited their capabilities. Specialized graphics-based languages for scientists is one area that shows significant promise. The BION™ contains a biologist-specific language that allows a scientist to communicate effectively with the system without learning arbitrary computer languages.

The future of scientific workstations holds great promise for tightening the link between human and computer. There is an increasing need to more tightly couple human problem-solving methods with software systems. Recent research in artificial intelligence has produced the concept of the knowledge-based system to provide such a link. Both factual and procedural knowledge, along with the strategies for reasoning with the knowledge, are represented in a manner natural to the human expert. Relatively simple reasoning methods, combined with large amounts of expert knowledge, have succeeded in providing expert or near-expert performance in areas ranging from molecular structure elucidation to antibiotic prescription to computer design and repair. Significantly, the user of a knowledge-based system is able to ask how a conclusion was reached and to receive an explanation which delineates specific facts, judgments, and associations used to reach a particular conclusion. This allows experts to "debug" knowledge bases interactively, and the systems to begin to serve as knowledgeable problem-solving assistants.

New computer architectures have been developed that support knowledge-based systems. Examples of systems currently available on the market include the XEROX 1100™ series of Scientific Workstations and the Symbolics, Inc. 3600™ Workstation. These systems combine facilities for carrying out artificial intelligence research with sophisticated interactive graphics.

The Future of Computers in Biotechnology

Biotechnology and computing are both changing rapidly. Computing hardware will continue to get less expensive, smaller, and more powerful; software will continue to get more expensive as the problems become harder. Artificial intelligence systems will start to play increasingly important roles in managing the most precious component of a high technology enterprise, the expertise of its scientists and engineers. Some potential emerging applications are discussed in the final paragraphs of this article.

The Intelligent Encyclopedia. One class of applications for knowledge-based systems is the intelligent encyclopedia, which allows a knowledge base to become an advisory system, or can provide a facility to "play out" the consequences of a strategy. Figure 2 is a display from an intelligent encyclopedia for the berries of California, run on the XEROX 1100 Scientific Workstation. BERRY allows the user to either view information about berries and their habitats or to "activate" the knowledge and have the system become an expert advisor on a particular berry. BERRY was constructed using the IntelliGenetics KEE™ system. KEE, the Knowledge Engineering Environment, provides a software facility for building intelligent encyclopedias or other knowledge-based systems.

Intelligent encyclopedias have also been applied to biotechnology. One example system built at IntelliGenetics provides a simulation environment for cloning experiments. A variety of DNA structures and proteins are represented in terms of attributes relevant to experiment planning; the representation of these structures can be easily changed by the user. The knowledge base contains sets of rules, described directly by a molecular biologist in close to "biological English," for generating restriction maps and simulating recombinant DNA experiment steps. The user can browse through the knowledge base of structures and strategies, and new structures can be generated by carrying out a simulated experiment. Figure 3 is a display from the knowledge base and a restriction map generated by carrying out such an experiment.

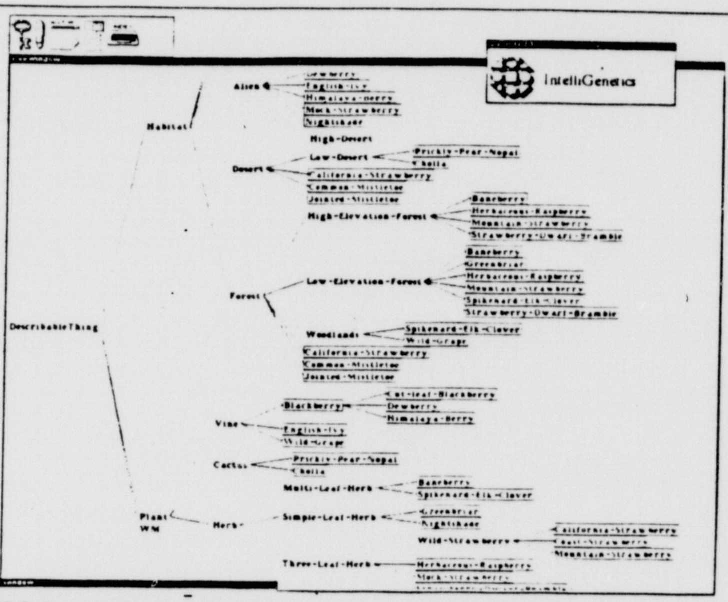

FIGURE 2. An example of a printout generated by an intelligent encyclopedia for the berries of California.

Design Aids For Scale-up. The ability of fermentation engineers to apply technical knowledge to problems in process modeling and process scale-up is of fundamental importance to the development of the biotechnology industry. Biochemical processes are immensely complex, involving theoretical and empirical information from microbiology, molecular genetics, and biochemical engineering. The crucial empirical knowledge tends to be judgmental, based on years of experience. Heuristic knowledge, typified by good judgment or good guessing, is an integral part of expert performance in fermentation development and control. Expertise

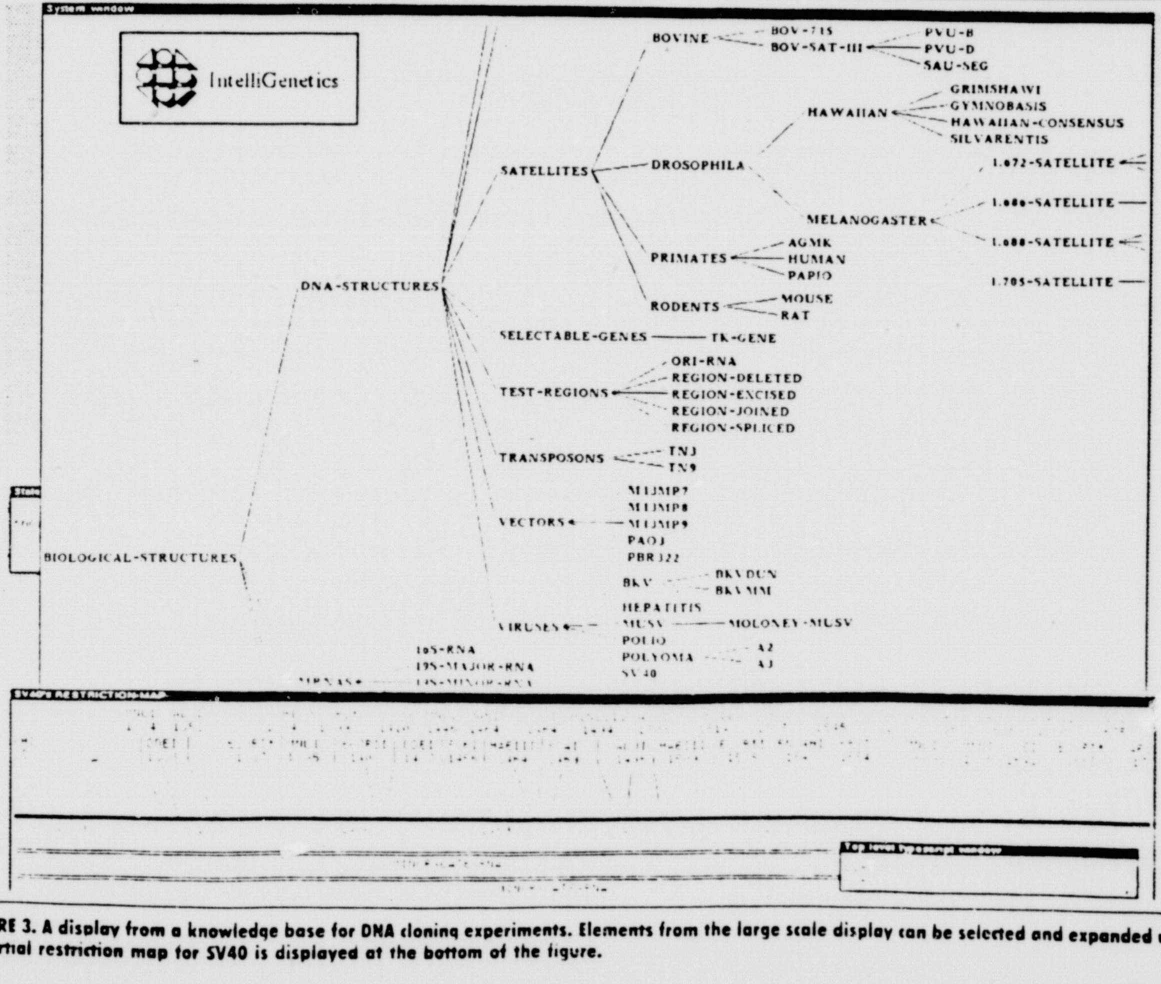

FIGURE 3. A display from a knowledge base for DNA cloning experiments. Elements from the large scale display can be selected and expanded upon. A partial restriction map for SV40 is displayed at the bottom of the figure.