III.A.2. Collaborative Research

Despite our budgetary restrictions, we have been able to make some progress in our Collaborative Research program. The most important step we have taken with respect to our program was to modify slightly our criteria for membership in Class II, our category for collaborative researchers. When BIONET was first established, the intent of Class II was for persons doing substantial development work on BIONET. We have decided this requirement is too strict for the following reasons:

- There are many programs of potential value to the BIONET community that have already been developed on other systems. Developers of these programs are willing to contribute them and do the work necessary for them to run on the 2060. Giving such contributors, selected for their community spirit, demand for their software and its complementary nature to other software available, Class II status seems to us eminently reasonable.

- Given our limited staff, it makes more sense to support a larger number of contributors who face only program compatibility problems, rather than a smaller number of developers each of whom might need substantial staff support. In any case, we have received several reasonable proposals from contributors and very few from developers.

For these reasons, we are entertaining, and have begun reviewing and accepting, proposals for Class II access from those who wish to contribute software to BIONET.

A second important step has been taken through the establishment of joint accounts at other molecular biology computing resources. We can now communicate through electronic mail with MBCRR, GenBank, and the PIR (see T. Smith, W. Goad and W. Barker in the list below). When our ARPANET connection (see Subsection III.A.3) has been established, communication and file transfer will be much easier, and we look toward direct exchange of programs and data over the network.

The following is a summary of our Class II community as of December, 1985. As of this date, this community has used about 8050 cpu minutes of computer time, and 1450 connect hours to BIONET. These figures represent about 6% and 5%, respectively, of the total BIONET Class I-III use of the system. These figures are indicative of the facts that several collaborators have just been accepted and have not yet contributed their software, and the others have primarily contributed their software or data and have required little development time.

M. Kanehisa/NIH. Dr. Kanehisa has contributed his IDEAS (Integrated Database and Extended Analysis System for nucleic acids and proteins) software. This suite of nine programs is a partial implementation of his VAX/VMS version of IDEAS, and contains eight programs for homology searches, three of which allow rapid database search, and an RNA secondary structure folding program. He installed these programs on the DEC-2060 with our help, recompiled all the software, and pointed all programs to our standard set of database directories and files. He has posted a message describing the availability of the programs on the BIONET-NEWS bulletin board; the bulletin has subsequently been moved to the CONTRIBUTED-SOFTWARE bulletin board. This bulletin has been followed by a review by BIONET staff members Drs. Azhir, Brutlag and Kedes, posted to the same bulletin board. This review pointed out some strengths and limitations of the programs, and contained some helpful hints on their use. Our records show that one or another of his individual programs have been accessed about 170 times since they were made available in April, 1985. This represents significant use by the community.

M. Zuker/NRC, Canada. Dr. Zuker contributed his program BIOFLD for RNA secondary structure folding to BIONET in March, 1985. He announced its availability through a message to the BIONET-NEWS bulletin board; this bulletin has subsequently been moved to CONTRIBUTED-SOFTWARE. R. Miller of W. Robinson's group at Stanford volunteered to review this program for the community. He has since posted a series of bulletins giving information on use of BIOFLD, suggestions on use of parameters and observations about the behavior of the program. BIOFLD has been accessed over 170 times by the BIONET community, representing significant use. Recently, Dr. Zuker has announced a PC version of BIOFLD to the BIONET community, and provided some preliminary documentation and instructions on obtaining a copy of the program.

D. Brutlag/Stanford. Dr. Brutlag and his student, B. Siegel, have begun a project to extend a program called MULTAN, for MULTiple nucleotide sequence ANalysis. This program was developed originally by B. Bains at Stanford, and will be available to the BIONET community. New developments will include translation into a more portable language, exploration of additional heuristics to improve upon initial selection of consensus sequence, improvement in the program's accuracy and application to analysis of polypeptide sequences.

H. Ginsburg/Minnesota. Dr. Ginsburg, in the laboratory of R. Dale, has begun a study of computer-based approaches to maintaining large collections of strains. This study will be carried out in the LISP programming language because of the utility of list representations to manipulations of strains and their genetic markers.

W. Pearson/Virginia. At his request, Dr. Pearson has licensed to IntelliGenetics, on a non-exclusive basis, the recent Lipman/Pearson "DFASTP" program for rapid protein homology searches. He has worked closely with us on producing a DEC-2060 version in the KCC C language compiler (see Paragraph III.A.5.d). The program is now running on the 2060 and will be released to BIONET after additional testing is performed. He has modified the program to read directly the original format of the Protein Database, maintained on the 2060 in the <NBRF> directory. We will continue working closely with Dr. Pearson as we extend DFASTP and produce the necessary documentation, but support for this version will be supplied by IntelliGenetics and BIONET.

D. Mount/Arizona. Dr. Mount originally proposed to make his PC software package available to the

BIONET community through down-loading of software to PC's. However, the slowness of file transfer programs such as Modem and KERMIT at 1200 baud make the time required prohibitively long. Recently, the Molecular Biology Computer Research Resource (MBCRR) at Dana Farber has begun floppy disk export of Mount's software. A bulletin to that effect was posted on BIONET-NEWS and has subsequently been moved to the CONTRIBUTED-SOFTWARE bulletin board. Thus, we do not expect to continue with our earlier plans to distribute the software directly from BIONET, but will direct requests to the MBCRR. We have written a letter of support to Dr. Mount for his application for a molecular biology computing resource. We feel that close collaboration among resources is essential to avoid duplication of effort.

T. Smith/Dana Farber. Dr. Smith, Director of the Molecular Biology Computer Research Resource, has been granted Class II access to BIONET by courtesy. This was done to facilitate cooperation and collaboration between BIONET and the MBCRR. Dr. Smith uses the bulletin board system on BIONET to announce availability of new software and data on the MBCRR system. For example, the Workshop on Problems in Genetic Sequence Analysis, scheduled for August, 1986, was announced to the BIONET community this way. Recently, the MBCRR has contributed to BIONET a version of the NBRF protein database restructured into functional categories. For example, all DNA-binding proteins, all immunoglobulins, and all cytochromes are grouped in individual files, and the files are in the standard format for use in the Core Library programs. We are currently testing the database prior to release of it to the BIONET community.

C. DeLisi/NIH. Dr. DeLisi has proposed contributing software for prediction of higher-order protein structures. Currently, the programs he feels are of most importance are still under development on his DEC-VAX facility.

G. Rose/Pennsylvania State. Dr. Rose has recently been accepted as a Class II collaborator and will be contributing software for protein secondary structure prediction.

C. Lawrence/NYS Dept. Health. Dr. Lawrence has recently been accepted and will be contributing software for statistical analysis of molecular biological data. He requires access to a library of statistical routines on BIONET, and the IMSL package of subroutines for statistics has been ordered for him and for other persons requiring access to these tools.

G. Stormo/Colorado. Dr. Stormo has recently been accepted and will be contributing software for quantitative sequence evaluation, analysis of binding sites, and sequence "landscapes" to display patterns of strings shared by two or more sequences. The last application represents another approach to solving the multiple sequence alignment problem.

W. Barker/NBRF,PIR. Dr. Barker is Director of the Protein Identification Resource (PIR), and has been given Class II status by courtesy. She represents our liaison with the PIR community.

W. Goad/Los Alamos. Dr. Goad heads the Los Alamos efforts related to the collection of nucleic acid sequence data for GenBank. He has been given Class II status by courtesy to foster communications with the GenBank Resource. He also collects sequences from BIONET submitted to him by electronic mail. He will contribute to BIONET programs for form-driven entry of sequences so that community members can submit their data in the correct format directly to GenBank.

R. Roberts/Cold Spring Harbor. Dr. Roberts is a member of our National Advisory Committee, so is grouped on the system in that category. However, he has spent a substantial amount of time working with BIONET on automated methods for updating his restriction enzyme database. Recently, he was able to transfer to us the latest version of this database in a format directly compatible with the Core Library software. Work remains to be done on automatic sending of messages about updates, and automatic logging of changes and testing of the new file, and we will assist him in completing these tasks. The goal is simple. We want BIONET scientists to have access to the latest data on restriction enzymes, rather than having to wait many months for its appearance on-line. Separately, Dr. Roberts is supplying a file of commercially-available enzymes, and we have already organized that into a form such that a user can programmatically select just those enzymes available from a selected supplier.

III.A.3. Core Research

Because of budgetary restrictions and the almost complete devotion of BIONET personnel and resources during the previous year to developing and consolidating the service, training, and collaborative components of the resource, Core Research has been limited to detailed planning of two major research goals for the next year of BIONET operations:

- **Hardware Text Searching Machines.** We are investigating specialized text searching hardware to optimize biological database searching;

- **BIONET Satellite Program.** We are investigating both hardware and software methods for the networking of BIONET with other regional, national, and international biologically-related computational resources.

III.A.3.a. Hardware Text Searching Machines

A common operation on BIONET involves the searching of one of the major nucleic acid or protein sequence databases for specific patterns of nucleotides or proteins. The Core Library of software has two programs that access these databases. The first is IFIND, which searches the database for sequence homologies using a specific query sequence. The second is QUEST, which is a sequence database search and retrieval program. QUEST uses a finite state machine that allows complex, often ambiguous patterns to be found in a database. Searches using either program against a large database such as the rapidly growing GenBank may require execution times ranging from cpu minutes to hours. Indeed, the growing use of batch jobs during nights and weekends (see Paragraph III.A.5.b) is a measure of the time required. Such searches represent a major use of cpu time on BIONET. Anything that can be done to reduce this time is not only scientifically interesting, it is essential in freeing up time for other scientists to perform their computations.

Recent hardware developments have led us to believe that we can vastly decrease the search time for complex patterns in QUEST. Such hardware may also increase the speed of the first phases of IFIND searches, and this application will be pursued after QUEST. One device, the Fast Data Finder (FDF) produced by TRW, Inc., can pass an entire database as one long character string through pattern matching hardware at a rate between 7 and 9 million characters per second. The databases are stored on a Fujitsu 2350 hard disk (474 Megabyte unformatted) driven by a Concept 21 disk controller which allows the formation of a very rapid data stream by interleaving data from several disk reading heads on the Fujitsu simultaneously. This multiplies the fundamental disk streaming rate from 1 to 1.5 megabytes per second per head, up to 7 to 9 megabytes per second, the limit of the FDF hardware. Transient rates above 10 megabytes per second are buffered in cache memory. The implications of these speeds are profound. For example, the GenBank database of nucleic acid sequences is now 12-14 Mbytes, including all comments. The FDF is capable of searching this database in 1.5 - 2 seconds.

The pattern to be found is stored in a series of cells in the FDF, one character per cell, and the data stream is passed through this series of cells. As the stream is passed from cell to cell through the FDF it reports a hit on the target when the pattern in each cell matches. The minimum number of cells (we are proposing initially 1,000 cells with the ability to upgrade to 10,000 cells in one year) would allow a maximum target size of 1000 characters. Much of the standard QUEST search key syntax (strings, ranges, fixed and variable length don't cares, Boolean relations etc.) is already built into the FDF hardware so that a straightforward translation of QUEST keys to FDF syntax is possible. We are proposing that TRW provide us with translations from our current pattern matching language into their syntax and also provide us with access to their Programmer Interface Language for interacting with the FDF. This will allow us to emulate QUEST in the easiest fashion.

The FDF has several advantages over the current QUEST program. First, the cells in the pattern matching hardware can be subdivided so as to search for several patterns simultaneously (maximum 248 patterns and each pattern utilizes a minimum of 24 cells although the patterns themselves may be smaller than this). Secondly, the FDF also allows up to seven mismatches within a defined character string within the pattern. These abilities to search for many patterns simultaneously and to permit mismatches in strings will allow the future development of DNA sequence alignment algorithms including rapid searches for homologies that including indefinite insertion/deletion gaps in addition to mismatches. For this later important application (after one year of use) we will need additional pattern matching cells, preferably near the hardware limit of 10,000. A further important property of the FDF is the ability to report regions of high density of specific sequence patterns. This has long been a major aim for QUEST development.

We intend to use our standard QUEST program on the BIONET DEC 2060 as the interface for users to prepare their search keys and to specify the database to be searched. We would also like the physical interface between the FDF (currently integrated with a SUN workstation) and the DEC 2060 to be flexible enough so that the FDF could be driven by identical software running on a VAX or on a SUN, the two other major machines that run QUEST. The simplest solution to this would be for the FDF to receive patterns and return results via a SUN based Ethernet connection.

The eventual goal is for the QUEST program to recognize when a proposed search will take more than a few moments of elapsed time and then ship the request over the Ethernet to the FDF hardware. The results of the search will be passed back to the QUEST program, so that the scientist using QUEST need make no special provision for long searches.

III.A.3.b. BIONET Satellite Program

We have begun the BIONET Satellite program in earnest. This program has the goal of distributing the BIONET Resource among computers throughout the academic community, while at the same time establishing better communication links among BIONET, its Satellites and other computing resources in molecular biology. Descriptions of the program with a more detailed statement of goals and objectives can be found in Appendix VI. As can be seen from the Appendix, the actual software license is a business arrangement between the Satellite institution and IntelliGenetics. BIONET's responsibility is to forge the communication links to ensure that scientists can communicate easily with one another.

We have previously described the initial, collaborative arrangements established with other resources in Subsection III.A.2. This is the first step toward the goal of linking the Resources. We currently have a Satellite established at the Salk Institute, and will soon establish two others, one at the US Department of Agriculture, the other at Fort Dietrick (US Army RIID).

We are following two approaches to communication with other facilities, ARPANET and a phone line based network that we are simply calling the BIONET Network for the moment.

ARPANET. BIONET has arranged Internet access to the ARPANET through a DARPA-funded project with IntelliCorp. In exchange for our assistance with the mechanics of the connection to ARPANET, BIONET will be able to make use of this connection for communications, especially electronic mail.

DARPA approved the IntelliCorp connection in October, 1985. We expect our connection to be operational in April, 1986, following the necessary lead times for the leased line provided by DARPA according to government procedures. Network services available through ARPANET include file transfer, mail, virtual terminal service, and others.

Since there are mail gateways from the ARPANET to other communications networks, this connection will do much to expand BIONET's reach. Most notably, mail interchange will be possible both with BITNET which includes EARN in Europe and with the NSF-originated CSNET. BITNET/EARN was undertaken collaboratively by a number of Universities with some help from IBM.

Additionally, since the ARPANET uses the TCP/IP internetwork protocols, a great many other networks with gateways to ARPANET will be fully accessible as well. These include the MILNET and local area networks at many major universities and research centers around the US and even in some foreign countries.

BIONET's central DEC-2060 resource will need to be connected to an IntelliCorp local area network which will in turn be a part of the Internet which includes ARPANET. This will mean that we will have to license software for the TCP/IP protocols for use on the 2060, and obtain an Ethernet interface to the local area network. At the same time, the IntelliCorp DARPA contract is purchasing the necessary gateway which will connect the IntelliCorp network to the leased line provided to an ARPANET network node processor, or IMP. The bandwidth of the ARPANET connection will be 56 kilobits per second, which will of course be shared by BIONET with the IntelliCorp DARPA users.

BIONET Network. As in the case of BITNET and CSNET, the ARPANET will form only a part of the communications backbone for the BIONET Network. The anticipated BIONET Network sites, or satellites, will vary in size and funding and an economical communications option is needed. We are currently examining options for hardware and software to provide this service.

We anticipate that asynchronous dial-up modems will be used to provide the economical link. As CSNET-RELAY does in CSNET, the BIONET central DEC-2060 resource will serve as the relay host for communication between BIONET's network sites. Most of the BIONET satellites are expected to be some model of the DEC VAX computer. BIONET has no-cost access to a MicroVAX II at IntelliGenetics and will develop the mechanism for mail exchange on this computer. We may wish to add a cache buffer memory to the DEC-20 front-end processor in order to increase the throughput possible for such communication.

III.A.4. BIONET Training Program

III.A.4.a. A Brief Review

The training program for BIONET has been severely restricted this year due to the budget cuts mentioned previously. However, we have been able to perform some trainings and demonstrate the use of BIONET at several national and regional meeting of molecular biologists. The presence of BIONET at meetings is not training in a formal sense, but there were many opportunities to answer specific questions and demonstrate use of BIONET for specific problems. These meetings also provide opportunities to inform potential BIONET applicants about the Resource.

The following summarizes our previous activities and those planned prior to the end of the current grant year.

- **FASEB Meeting**. The Federation of American Societies for Experimental Biology meeting was held April 22-25, 1985, in Anaheim. At this meeting we made a formal presentation about the BIONET Resource, in a *Workshop on International Genetic Sequence Resources*. In addition, we participated in a booth, jointly sponsored by IntelliGenetics and BIONET.

- **Rutgers/Waksman Institute Workshop**. A workshop, entitled *INTRODUCTION TO BIONET: A National Computer Resource for Molecular Biology* was held under the auspices of the Waksman Institute of Microbiology, at the Piscataway campus of Rutgers University, June 17-19, 1985. There were two parts to the Workshop, a one-day lecture program on June 17, attended by 79 persons, followed by two additional days for 23 people, all of whom attended the first day. The program for this Workshop is shown in Appendix V. The two-day session allowed all attendees access to terminals connected to a DEC-2060 machine at Rutgers running the Core Library software and emulating the BIONET bulletin board and electronic mail systems. The reports from all attendees on their reactions to the training were extremely positive. All left feeling they know much more about the use of computers in molecular biology in general, and the use of BIONET in particular. The most frequent negative comment was that there was too much material covered in the one-day session.

- **NATURE Meeting**. The *NATURE* meeting entitled *Update in Molecular Biology* was held October 7-9, 1985 in San Francisco. IntelliGenetics and BIONET jointly sponsored a booth at the show.

- **BIOTECH '85**. The BIOTECH '85 International Conference and Exhibition was held October 21-23, 1985 at the Washington Convention Center, Washington, DC. BIONET and IntelliGenetics jointly sponsored a booth at the exhibition.

- **International Congress on Computers in Biotechnology**. This congress will be held January 30-31, 1986 at the Baltimore Convention Center. A talk will be presented on the BIONET Resource in a session titled "Systems and Resources". BIONET information will be available at the IntelliGenetics booth set up in conjunction with other, overlapping conferences sponsored this same week at the Convention Center.

- **Miami Mid-Winter Symposia**. BIONET will sponsor a booth at the Mid-Winter Symposia in Miami, February 3-7, 1986. We are arranging for two training sessions at the meeting, organized around new training materials discussed below.

III.A.4.b. Some Lessons Learned

The trainings at Stanford late in the first year of our grant, the training at Rutgers/Waksman and our experience in assisting the scientific community at trade shows and in our extensive scientific consulting all lead to the same conclusion. People, especially those unfamiliar with computers, get very little out of lectures on use of software. Without the ability to use a system under careful guidance, the amount of information transferred is only slightly above zero. There must be terminals and/or PC's, at least one per two trainees, access to the BIONET software and communication facilities if not the actual computer itself, and carefully chosen examples to illustrate use of both system and application software. Despite our efforts to write documentation for the new user, it is clear that available documentation and training manuals are useful only after a person has mastered some basic techniques.

III.A.4.c. A New Strategy

We are going to develop a new training program, built around examples of application of our software to problems described in the language of molecular biology. This will differ substantially from our current materials, which are focused on specific programs and what they will do, rather than on a specific problem and how to solve it. Our experience has shown us that the following kinds of topics would cover the questions asked most often (these examples are part of a bulletin that was sent to potential participants at Miami):

- BIONET: FACILITIES AND COMMUNICATIONS.
    - What programs and features are available to BIONET users: descriptions of what each is typically used for and how you can access them
    - How to master UNINET
    - How to find important information of the bulletin boards
    - How to keep your directory within allocation
    - How to send electronic mail--including how to find out who else is on BIONET
    - How to make your backspace key work
- ENTERING AND EDITING DNA AND PROTEIN SEQUENCES
    - Using the screen-oriented editors (ESEQ); deciding what type of "terminal" you are for GENED; how to move the cursor in the editor
    - How and when to use ambiguity codes
    - Entering proteins by three-letter codes
    - Creating subsequences out of known sequences
    - Selecting and saving a sequence from the database for your own use

- GENERATING RESTRICTION MAPS--FINDING RESTRICTION ENZYME CUT SITES
  - Listing all or a subset of restriction enzyme cut sites of your sequence
  - Generating restriction maps from fragment size or mobility data
  - Generating restriction maps of a given sequence
  - Creating and using an individualized restriction enzyme list
- CONSTRUCTING VECTORS
  - Locating and using existing maps of common vectors
  - Cleavage and recombination of fragments
  - Generating a cloning vector restriction map
  - Excising fragments to customize recombinant plasmids
  - Testing directional cloning and insertional inactivation in cloning vectors
- ASSEMBLING SEQUENCES TO GENERATE A CONSENSUS SEQUENCE
  - Entering gel sequence information
  - Automatically merging together data from multiple gels
  - Editing consensus sequence--how to propagate changes through to constituent gels
  - Error checking and sequence comparison
  - Handling of both dideoxy and chemical sequencing data
- SEARCHING and ALIGNMENTS
  - How to find out if your sequence is in the database
  - Comparison of your sequence vs. the entire database
  - Comparison of your sequence vs. taxonomic or some functionally similar partitions of the database
  - Explanation of indirect files
  - How to search for sequences with key words or literature references
  - What alignment methods are available, and which to use when
- OTHER COMMON ANALYSES
  - Searching for optimal regions to design probes
  - Reverse translation

  - Hydropathicity plots (and what each method's graphs mean)
  - Secondary structure prediction
  - Calculating amino acid composition
  - Translation
  - Searching for dyad symmetries
  - Locating internal repeats
  - Calculating base composition

- FILE TRANSFER
  - How to get your PC to act like a terminal
  - How to get data to and from BIONET

III.A.5. Resource Facilities

There have been several changes in the management of and personnel assigned to the BIONET computer facilities. These changes are summarized in Section III.C, Administrative Changes. The present section is devoted to a description of the current facilities and summary statistics on use of the Resource. The statistics cover the twelve months since our last Annual Report, 12/84 - 11/85.

III.A.5.a. Computer Hardware and Telecommunication Networks

Hardware. The BIONET Central Resource Machine is a Digital Equipment Corporation 2060 computer. The configuration was augmented this year to include an additional RP07 disk drive. Rather than simply providing additional disk space, this drive allows us a fallback in the event of the failure of one of the primary RP07 drives. (This happened during the month of October, 1985, and the existence of the additional RP07 did in fact greatly reduce the necessary downtime.) The primary drives are combined into a single disk structure and must both be functional in order for the system to run. In addition, the third RP07 is used as an additional storage place for files which are not essential in a short-term fall-back operation.

The hardware configuration is as follows:

```
KL10-E Model R Processor:

      2 MF20/MG20 Memory controllers
   2 MW MG20 Memory
 .75 MW MF20 Memory
        MCA20 Cache Buffer Memory
      2 RH20 Massbus Channels
```

Console and Front End Processor:

PDP-11/40 CPU, 32 KW 16 bit memory
RX02 Dual floppy disk drives
8 DH11 Terminal interfaces   8 * 16 TTY lines each = 128 lines
RH11 Massbus Channel
LP20 Line printer interface

DN20 Front End Processor:

PDP-11/34 CPU, 128 KW 16 bit memory
DMR11 Network interface

Peripherals:

3 RP07 disk drives   111MW each
RP06 disk drive   39MW
372 MW Total disk storage
TU78 1600/6250-BPI tape drive
LP26 600 LPM Line printer
Imagen Imprint-8/300 Laser Printer

Disk space (data storage)

Public structure (PS:) disk space use on the 2060 is dynamic. The following snapshot is representative of typical usage, and is taken from December 1985.

Total disk space	433,000	(pages--222 million words)
Overhead/Common	<148,000>	(Core, System and System Support Libraries)
Swapping Space	< 25,000>	
File system Overhead	< 70,000>	(Directories and index pages)

	190,000	
BIONET Allocation	95,000	(Half of the available space)
Bionet Usage 12/85	< 53,000>	

Unused space	42,000	(Available for BIONET growth)

Note that file system overhead varies greatly depending on the size of the files involved. Since BIONET users have many small files, BIONET growth may increase file system overhead, altering the above distribution.

Terminal Lines

Because the usage of a particular terminal line varies greatly, and because many BIONET users share a single line in succession, there was in the past an imbalance in the allocation to BIONET of terminal lines. However, with the departure of the IntelliCorp KSD users from the system (see Section III.C), additional terminal lines were freed for BIONET. These are not regularly needed by BIONET at this time, but may be used intermittently or for growth. Current system terminal line distribution is as follows:

Total lines	128	
Overhead	< 10>	(Shared devices, BCRG staff)

	118	
Allocated BIONET	59	(Half of the available lines)
BIONET Users	< 18>	(Public Data Network, Local Dial-Ups)
BIONET Staff	< 6>	

Unused lines	35	(Available for BIONET growth, temporary use for trainings, replacement of a bad line before it is repaired)

Public Data Network Connection. BIONET is accessed principally over the UNINET Public Data Network. An X.25 PAD (packet assembler/disassembler) is located on-site. This is known as the Host PAD, or HPAD. It provides individual terminal ports which are cross-connected to those on the DEC-20. The Uninet trunk line operates at 9600 baud synchronously, and the PAD converts this into up to 16 asynchronous ports whose speed is typically 1200 baud. A handshaking protocol is employed to smooth over bursts of data during the multiplexing.

UNINET we originally chosen as a replacement for Telenet because of its better response time and its lower cost. The lower cost was achieved through a very favorable fixed price per port arrangement that we negotiated with UNINET. Currently 12 UNINET host ports are used by BIONET, and usage is monitored carefully in the event more are needed. The ports are accessed in sequence, with those higher in the sequence not being used while any lower port is free. The number of connect hours per month drops off after the first 6 ports. The usage on these first 6 ports therefore represents many more sessions than does the usage of ports 7 through 12. Our monitoring of the port use also has revealed that it would be cheaper for BIONET to lease the higher-numbered ports on a use, or traffic, basis. We currently are leasing 8 ports fixed, 4 on traffic, and will change this distribution as required for the lowest possible cost.

We have been examining the replacement of the UNINET-supplied leased HPAD with a BIONET owned HPAD. The consideration is the savings of lease charges while maintaining adequate reliability. We plan to make this replacement before the end of the current grant year.

III.A.5.b. Summary Statistics on Machine Use

The cpu cycles of the DEC-2060 computer are allocated to the user community, including BIONET, by the system's class scheduler. This scheduler is given the percentage of the machine to allocate to each class of users. Any cycles not consumed by a given class ("windfall")are available to the rest of the user community. This method was chosen so that cpu cycles not consumed by one segment of the community could be used by other segments if needed, i.e., no cpu cycles are wasted if someone needs them.

The current percentage allocations ("pieslices") are shown in Figure III-1. As summarized in the figure, BIONET Class I (and III and IV) are allocated 30% of the machine, and Class II and staff 10%. The 20% overhead (system overhead, batch and computer staff and operations) is allocated one-half to BIONET, for a total of 50%. These allocations remain the same as last year. However, there are substantial changes to the other classes of users for reasons discussed in Section III.C. Note that the BATCH class is assigned 1% of the system during prime time. In off prime time, the percentage allocation is increased substantially in response to demands by the BIONET community.

The actual use of the machine by the BIONET community is now substantially greater than 50% of the total cpu cycles actually used. As an example, the percentage use of the machine for the month of October, 1985 is shown in Figure III-2. It is clear that BIONET is receiving more than its fair share of the cpu cycles. Note that BIONET scientists' use of BATCH is charged to the individual accounts by the accounting program. Thus, extensive use of BATCH shows up in this pie chart as BIONET Class I (or II) use, rather than in in the category BATCH Jobs.

The data for BIONET percentage of system use are plotted in histogram form in Figure III-3. This figure demonstrates that BIONET has consumed more than 50% of the total cpu cycles **used** (data on % of **available** are given below) on the 2060 since February, 1985, and is now consistently consuming 65 - 75% of the total cpu cycles used on the system.

In the following series of tables and figures, we provide further details on the actual use of the system by the BIONET community. Looking first at use of the system in prime time (8 AM - 8 PM, M-F, PST), data for cpu time and connect hours for the indicated segments of the community are given in Tables III-4 and III-5 by month, and totals. The cpu data in Table III-4 is also plotted in histogram form in Figure III-4. (The figures for the facilities group staff and overhead for November, 1985 are artificially low because the statistics were computed before Thanksgiving weekend, before the end of the month operator totals were added in.)

There are several important facts that can be determined from these data. Looking first at cpu time, and given that there are about 12,000 cpu minutes (total cpu minus 20% for overhead) **available** prime time in the average month for the entire system, BIONET (Users plus Staff) has been consuming well over 50% of available cycles. The category of BIONET Users (Classes I-III) compete for 30% of the machine. The class has consumed more than 30% of available cycles since March, 1985, and have thus been able to take advantage of considerable windfall.

Figure III-1: Pieslice Allocation of the DEC-2060 Computer

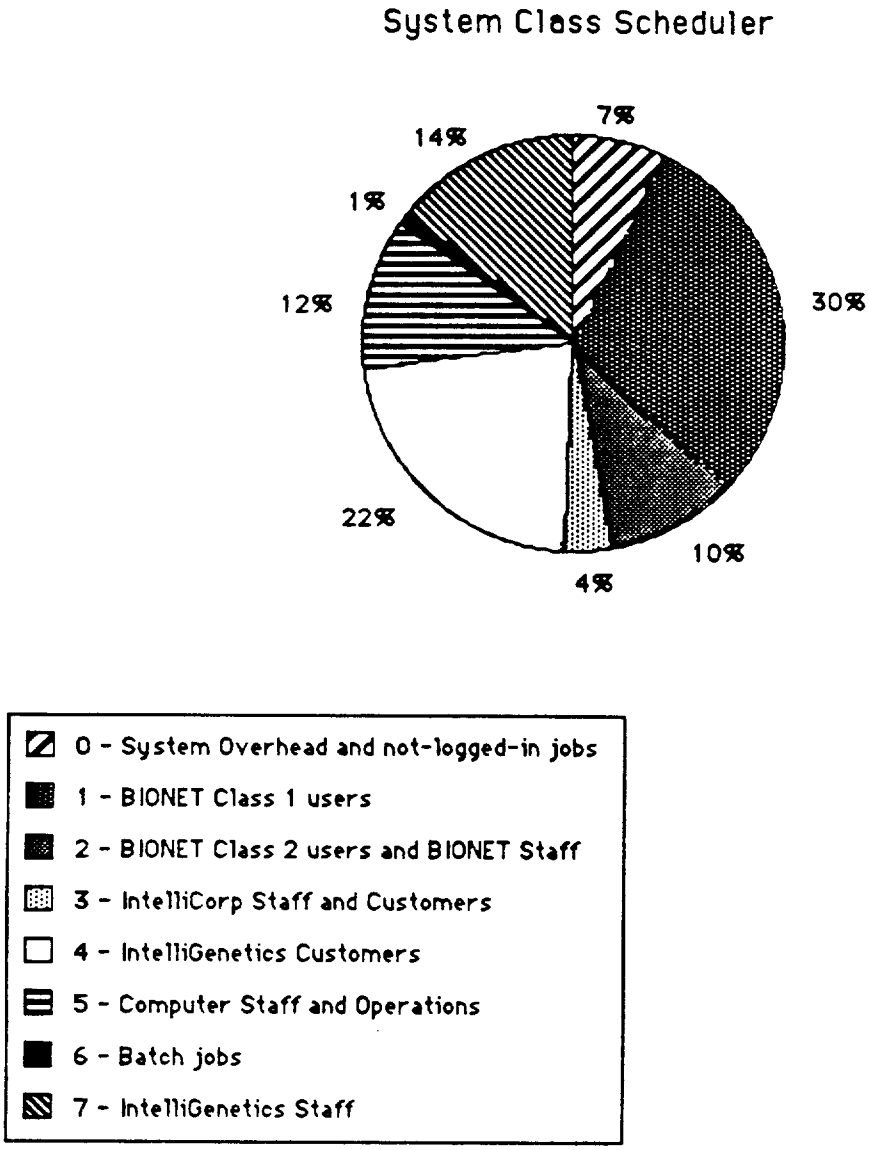


Figure III-2: Actual Use of the DEC-2060 for the Month of October, 1985

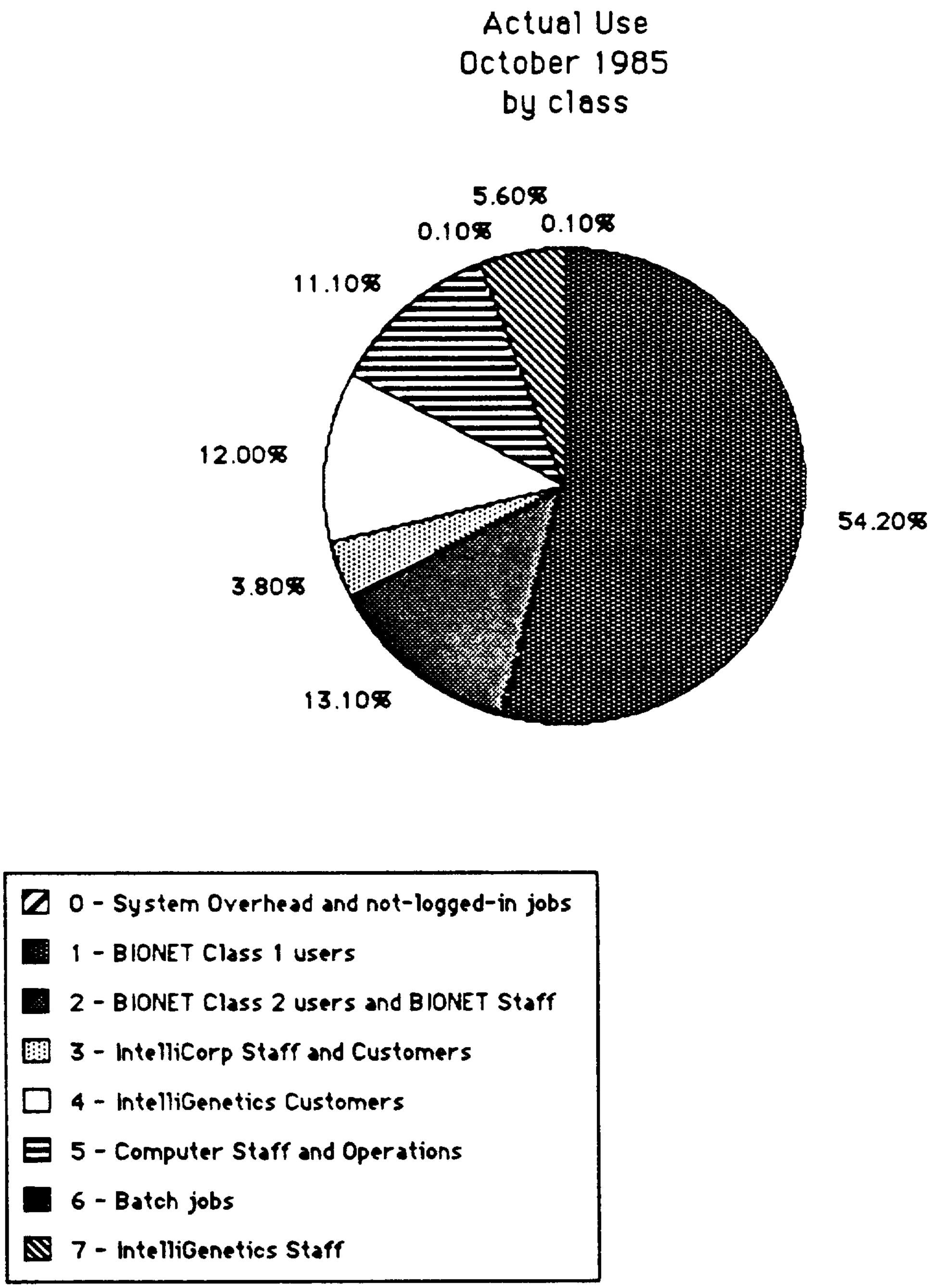
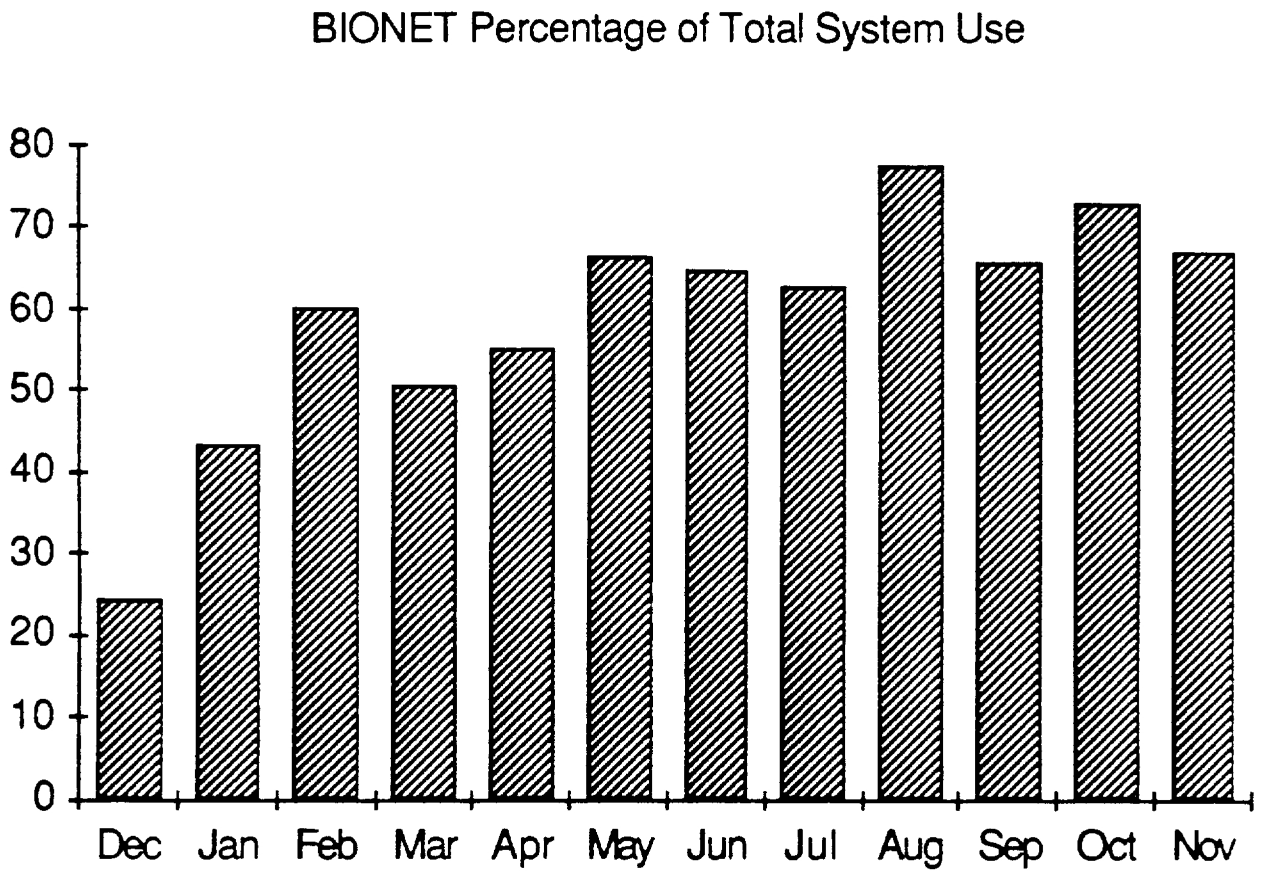


Figure III-3: BIONET's Percentage Use of the DEC-2060, 12/84 - 11/85



The total number of connect hours, prime time (Table III-5), for the category BIONET Users has remained in the range 1800 to 2200 since May, and the relationship between connect hours and cpu minutes remains relatively constant over those months.

The data for non-prime time (weekends and 8 PM - 8 AM M-F) are shown in Tables III-6 and III-7, and the data on cpu time are plotted in histogram form in Figure III-5. Particularly notable in these data are the dramatic increases in cpu time over the past year, especially in the last three months, due almost entirely to BIONET use. These increases are due primarily to the extensive use of overnight batch runs to perform time-consuming analyses involving database searches, using the IFIND homology and the QUEST database search and retrieval programs. Thus, the community has gravitated naturally toward off-hours use of these programs for such analyses.

Given that there are about 22,000 cpu minutes (total minus 20% overhead) available each month in non-prime time, BIONET (Users plus Staff) has recently been consuming more than 50% of the amount available. Given low use of the system by other classes in non-prime time, BIONET consumes most of the cpu cycles actually used during these times.

The data for total use of the Resource by BIONET are presented in Tables III-8 and III-9 and the total cpu time is summarized in Figure III-6. BIONET Users and Staff, since May of 1985, have consumed 40% or more of all the cpu cycles available on the system (total minus 20% overhead).

One important conclusion from all these data is that the Resource is rapidly approaching saturation. Certainly, during prime time, the system load is becoming a barrier to rapid computation. At this point, limitations on the number of access ports keep the load average under control by limiting the number of concurrent users. However, as we add additional telecommunication ports, we will quickly become limited by available cpu time.

Another important conclusion we have reached from these data is in regards to the effects of the subscription fees on use of the Resource. The total use by BIONET scientists (not including staff) increased steadily from November, 1984 through May, 1985. In the summer months of June through August, use leveled off, beginning before the subscription fee was announced, which we attribute to summer vacations more than any effect of subscription fees. Beginning in September, 1985, use increased steadily again to a level substantially above the months prior to initiation of the fee.

Summary data for use of our telecommunications network are presented in Figure III-7 by month for the past 12 months' use of the Telenet (until mid-July, 1985) and UNINET (beginning early July, 1985) networks. Three factors distort this Figure. First, the value for July is artificially high because we were running the two networks simultaneously and performing extensive tests on UNINET. Second, we noticed

Table III-4: BIONET Prime Time CPU Minutes

	BIONET Users (except staff)	BIONET staff	BCRG & System Overhead	Total BIONET Use
December	769	397	385	1551
January	2598	1054	579	4231
February	3368	1091	644	5103
March	4236	571	473	5280
April	5169	861	529	6559
May	6791	776	515	8082
June	5004	905	530	6439
July	5575	1094	564	7233
August	5132	1248	508	6888
September	4854	798	509	6161
October	6476	1330	455	8261
November	6135	473	88	6696
TOTAL	56107	10598	5779	72484

Table III-5: BIONET Prime Time Connect Hours

	BIONET Users (except staff)	BIONET staff	BCRG & System Overhead	Total BIONET Use
December	328	519	1218	2065
January	761	1164	1368	3293
February	1137	829	1340	3306
March	1206	638	347	2191
April	1452	764	1353	3569
May	2177	737	1473	4387
June	1908	577	1567	3690
July	2291	916	1661	4643
August	1846	700	1374	3767
September	1777	606	1585	3810
October	2101	763	1688	4537
November	2187	689	156	3032
TOTAL	19171	8902	15130	42290

Figure III-4: BIONET's Prime Time Use of the DEC-2060, 12/84 - 11/85

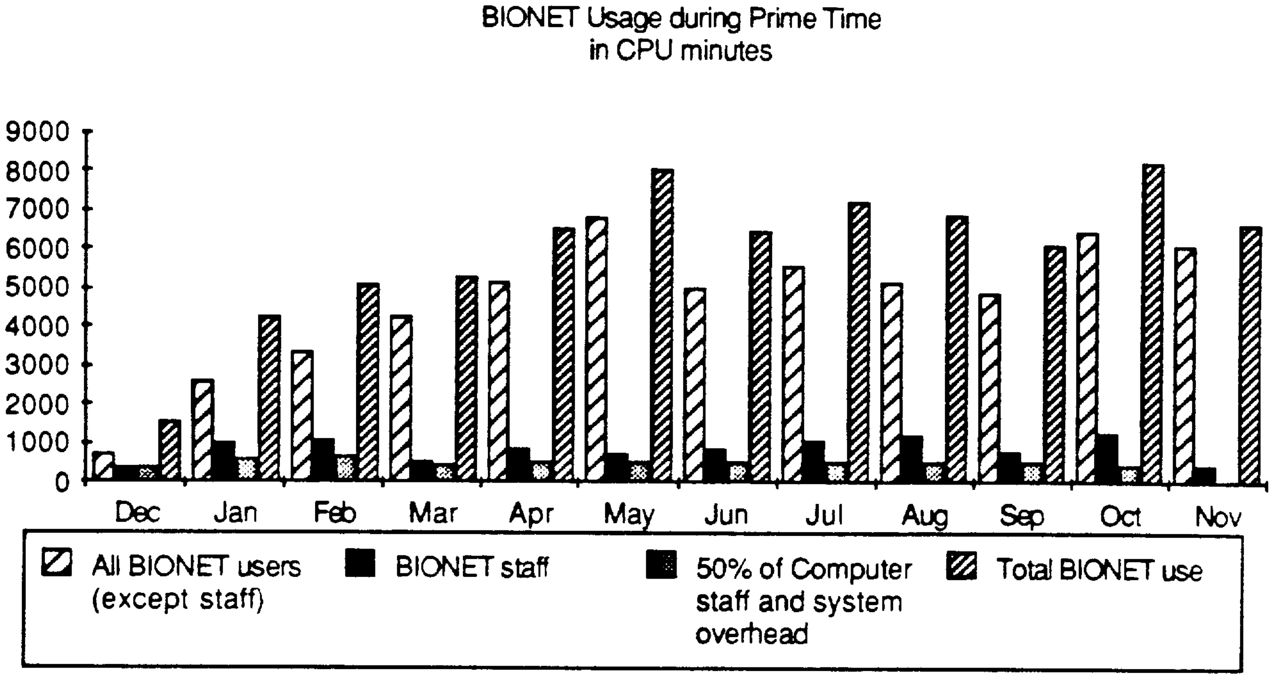


Table III-6: BIONET Non-Prime Time CPU Minutes

	BIONET Users (except staff)	BIONET staff	BCRG & System Overhead	Total BIONET Use
December	366	91	225	682
January	1673	128	826	2627
February	3848	357	159	4364
March	4169	26	404	4599
April	3386	356	1370	5112
May	6777	206	1300	8283
June	6567	1129	1415	9111
July	6956	850	1613	9419
August	5396	1244	1238	7878
September	7056	1192	876	9124
October	9553	1407	1103	12063
November	12326	111	86	12523
TOTAL	68073	7097	10615	85785

Table III-7: BIONET Non-Prime Time Connect Hours

	BIONET Users (except staff)	BIONET staff	BCRG & System Overhead	Total BIONET Use
December	117	159	1751	2027
January	420	145	1749	2314
February	562	208	1680	2450
March	697	121	142	960
April	601	149	1859	2609
May	949	221	2002	3172
June	1246	194	2210	3650
July	1197	230	2519	3946
August	887	192	1843	2922
September	1109	202	2590	3901
October	1213	190	2343	3746
November	1746	173	117	2036
TOTAL	10744	2184	20805	33733

Figure III-5: BIONET's Non-Prime Time Use of the DEC-2060, 12/84 - 11/85

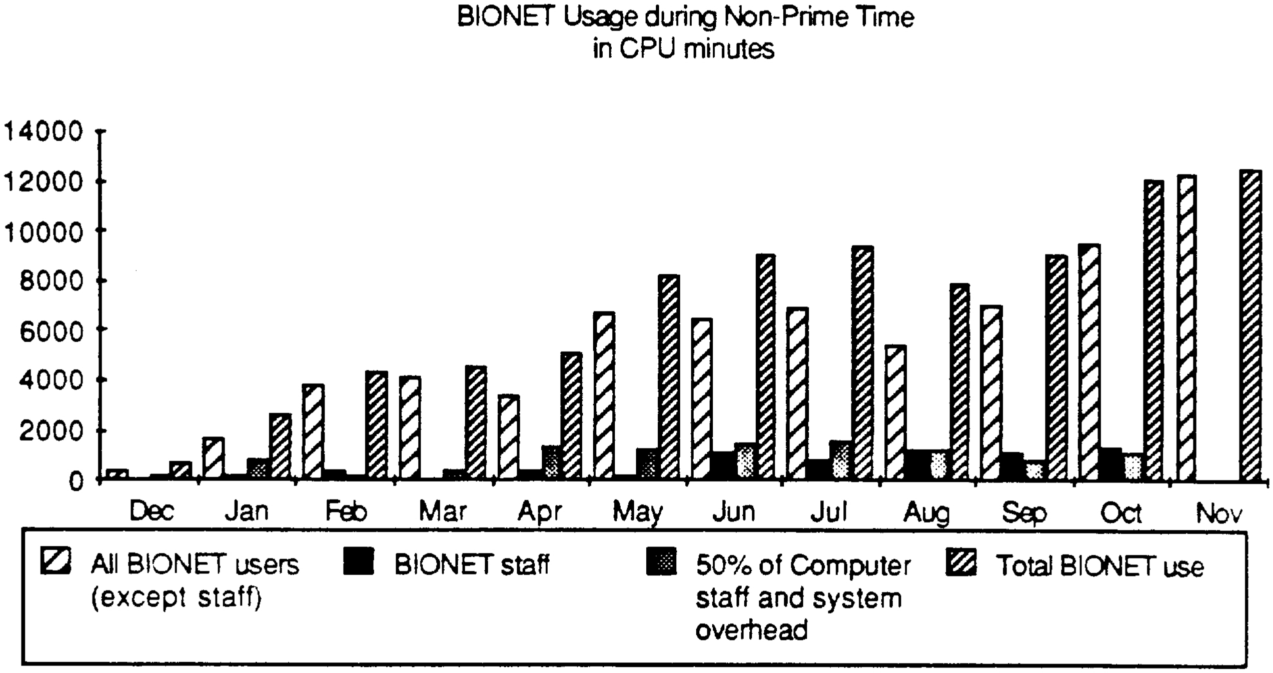


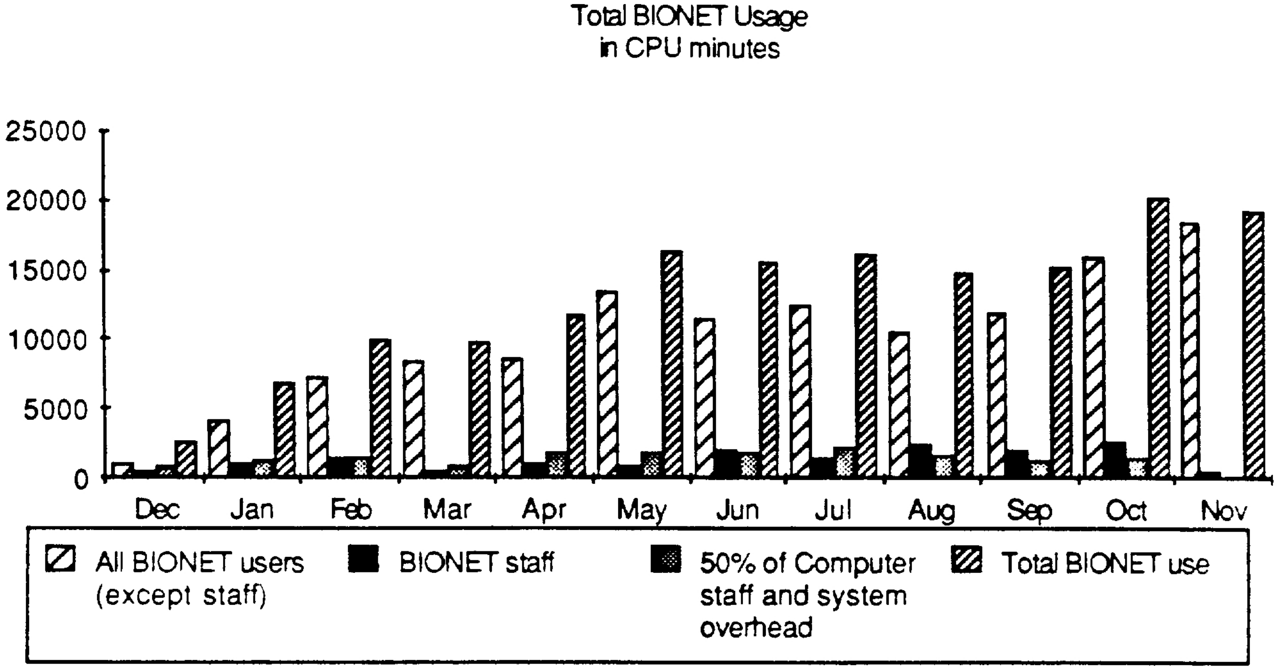
Table III-8: BIONET Total CPU Minutes

	BIONET Users (except staff)	BIONET staff	BCRG & System Overhead	Total BIONET Use
December	1136	489	1015	2640
January	4271	1182	1407	6860
February	7216	1449	1414	10079
March	8405	597	877	9879
April	8556	1217	1899	11672
May	13568	982	1816	16366
June	11571	2035	1946	15552
July	12531	1945	2178	16654
August	10528	2492	1747	14767
September	11911	1990	1386	15287
October	16029	2737	1559	20325
November	18462	585	174	19221
TOTAL	124184	17700	17418	159302

Table III-9: BIONET Total Connect Hours

	BIONET Users (except staff)	BIONET staff	BCRG & System Overhead	Total BIONET Use
December	445	678	2969	4092
January	1181	1309	3117	5607
February	1699	1037	3020	5756
March	1903	759	489	3151
April	2053	913	3212	6178
May	3126	958	3475	7559
June	3154	771	3777	7340
July	3488	1146	4180	8589
August	2733	892	3217	6689
September	2886	808	4175	7711
October	3314	953	4031	8283
November	3933	862	273	5068
TOTAL	29915	11086	35935	76023

Figure III-6: BIONET's Total Use of the DEC-2060, 12/84 - 11/85



that many users were leaving their terminals after completing their work without logging off BIONET, thereby tying up the network port and preventing other users from accessing that port. Therefore, we implemented an "idle zapper" which monitors the cpu use for each BIONET job, sends a warning message after 10 minutes of cpu idle time, and detaches the job after 5 more minutes of idle time, as a good compromise based on comments on the idea from the user community. Thus, an idle job can tie up a port for no longer than 15 minutes. The job is still available to the user, who can reattach to it and continue from where he or she left off. The zapper has been very effective in freeing up network ports. Third, the data for October, 1985 are artificially low because of UNINET network problems, which have since been resolved.

III.A.5.c. Computer Software - Core Library

Through our license agreement with IntelliGenetics, we have provided all Core Library software releases to the community. There have been two major releases so far this grant year, and another will occur at the end of January, 1986.

One important addition to the Core Library was requested by Dr. Yanofsky of our National Advisory Committee, the addition of the DIGITIZER program to the suite of software. Up until recently, access to the software to use a sonic digitizer for entry of gel data (restriction digests, sequencing ladders) has not been possible for BIONET scientists. We have made arrangements to modify the software license agreement with IntelliGenetics, and digitizer access is now possible. A bulletin to that effect has gone out to the community, and a small number of laboratories have purchased the necessary hardware to use DIGITIZER.

III.A.5.d. Computer Software - System Library

During the course of the year, the following additions have been made to the system support library described in last year's report.

Communication.

FINGER--Displays an information message or "plan" optionally provided by a user for other users advising them of travel itinerary or other contact information, and also displays the date the user in question was last on the BIONET system.

WHOIS--Directory lookup program for BIONET investigators. During the course of the year this utility was upgraded to have more generalized search capability and to permit searches of mixed case text. The WHOIS database of BIONET users was extended to include research titles for each PI, to enable other PI's to identify investigators with similar research interests.