II.B SUMMARY OF RESOURCE USAGE

The following data give an overview of the resource usage from May 1975 through April 1976. There are three sub-sections containing data respectively for 1) resource usage by community (AIM, Stanford, and system), 2) resource usage by project, and 3) Network usage data.

II.B.1 RELATIVE SYSTEM LOADING BY COMMUNITY

The SUMEX resource is divided, for administrative purposes, into 3 major communities: user projects based at the Stanford Medical School, user projects based outside of Stanford (national AIM projects), and systems development efforts. As defined in the resource management plan approved by BRP at the start of the project, the available resource will be divided between these communities as follows:

CPU Usage -	Stanford	40%
	AIM	40%
	Staff	20%
File Space -	Stanford	27,000 pages(*)
	AIM	27,000 pages
	Staff	13,500 pages

(*) One TENEX page is 512 36-bit words or 2560 text characters)

An additional allocation of approximately 30,000 pages serves system files including documentation, subsystems, monitor, etc.

The monthly usage of CPU and file space resources for each of these three communities relative to their respective aliquots is in the plots in Figure 8 and Figure 9. Our diurnal variations in loading have retained the same characteristics as previously, with a bimodal distribution reflecting the complementary loads from the east coast and the west coast.

Figure 8.
CPU USE BY COMMUNITY

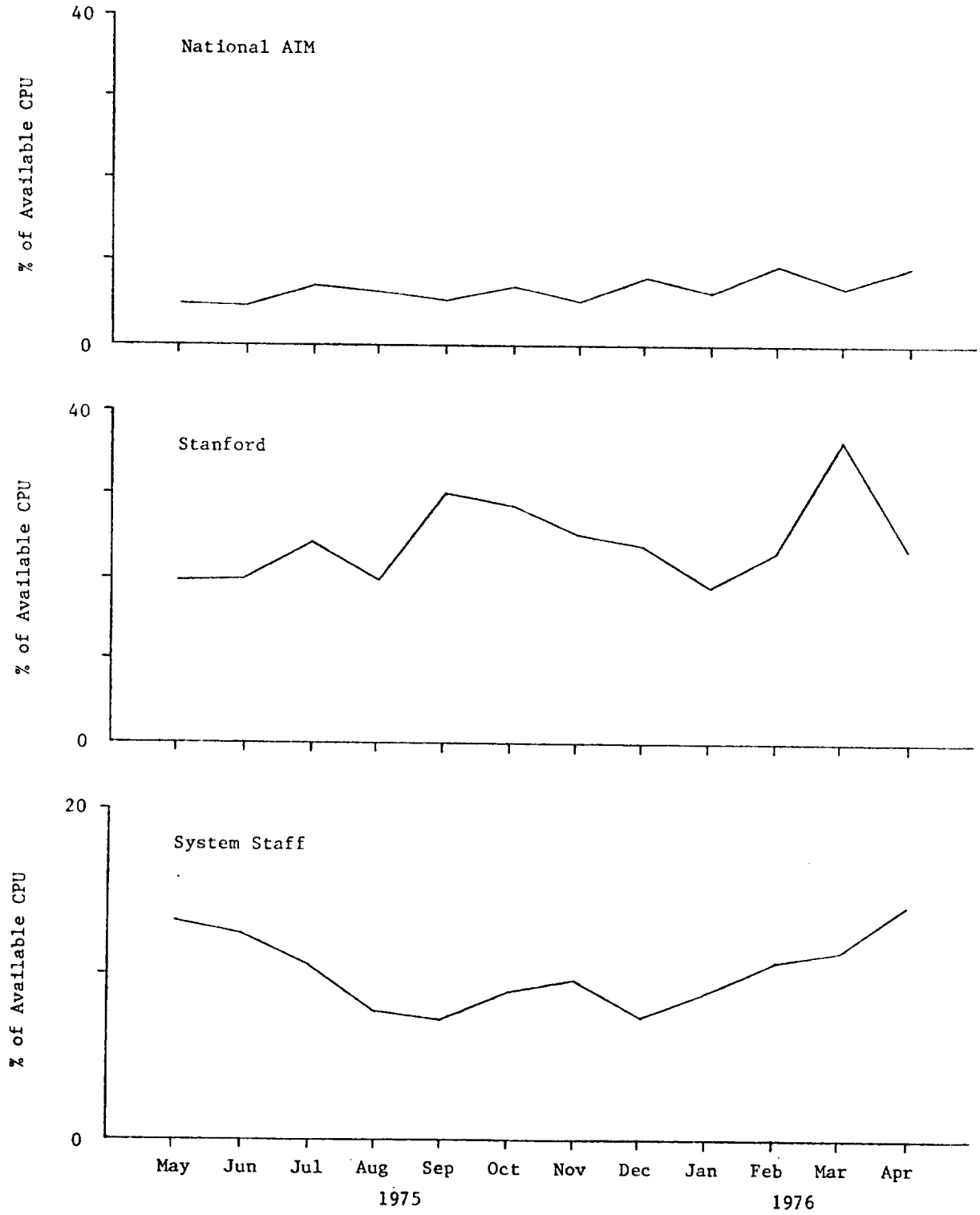
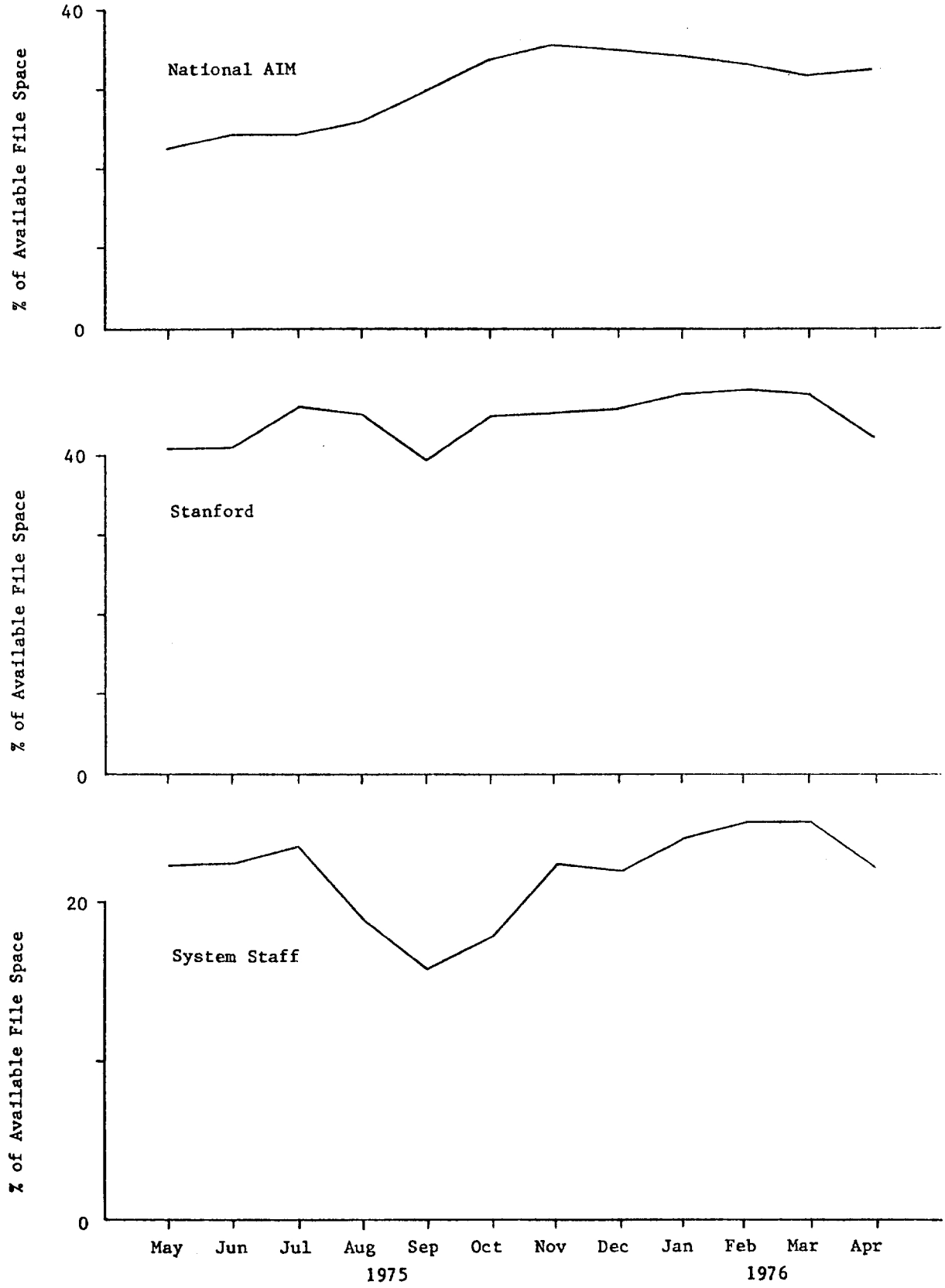


Figure 9.
FILE SPACE USE BY COMMUNITY



II.B.2 INDIVIDUAL PROJECT AND COMMUNITY USAGE

The table following shows average resource usage by project in the past grant year. The data displayed include a description of the operational funding sources (outside of SUMEX-supplied computing resources) for currently active projects, average monthly CPU consumption by project (Hours/month), average monthly terminal connect time by project (Hours/month), and average file space in use by project (Pages/month, 1 page = 512 computer words). Averages were computed for each project for the months between May 1975 and April 1976.

RESOURCE USE BY INDIVIDUAL PROJECT

STANFORD COMMUNITY	CPU (Hrs/mo)	CONNECT (Hrs/mo)	FILE SPACE (Pages/mo)
1) DENDRAL PROJECT "Resource Related Research Computers and Chemistry" NIH RR-00612 (3 yr award) $240,967 this year	68.41	1574	18280
2) MYCIN PROJECT "Computer-based Consult. in Clin. Therapeutics" HEW HSO-1544 (3 yr award) $163,965 this year	20.76	494	5959
3) PROTEIN STRUCT MODELING "Heuristic Comp. Applied to Prot. Crystallog." NSF DCR74-23461 (2 yrs.) $88,436 total	19.45	296	2452
4) PILOT PROJECTS (see reports in Sec IV.B.1)	14.12	433	3459
COMMUNITY TOTALS	122.74	2797	30150

NATIONAL AIM COMMUNITY			
1) SECS PROJECT "Chemical Synthesis" NIH proposal pending	10.32	196	3284
2) INTERNIST PROJECT (DIALOG) "Computer Model of Diagnostic Logic" HEW MB-00144 (3 yrs.) $167,168 last year	7.64	209	4705
3) Higher Mental Functions "Computer Models in Psychiatry and Psychother." NIH MH-27132 (2 yrs.) $67,000 this year	1.94	85	1299

4)	ACT PROJECT "Language Acquisition Modeling" NIMH $20,000 this year	2.77	55	559
5)	MISL PROJECT "Medical Information Systems Laboratory" HEW MB-00114 (3 yrs.) $248,793 this year	0.98	45	773
6)	RUTGERS PROJECT "Computers in Biomedicine" NIH RR-00643 (3 yrs.) $314,880 this year	12.17	446	8174
7)	AIM PILOT PROJECTS	0.27	9	66
8)	AIM Administration	1.88	76	2897
		---	---	---
	COMMUNITY TOTALS	37.97	1121	21757

SUMEX STAFF AND SYSTEM

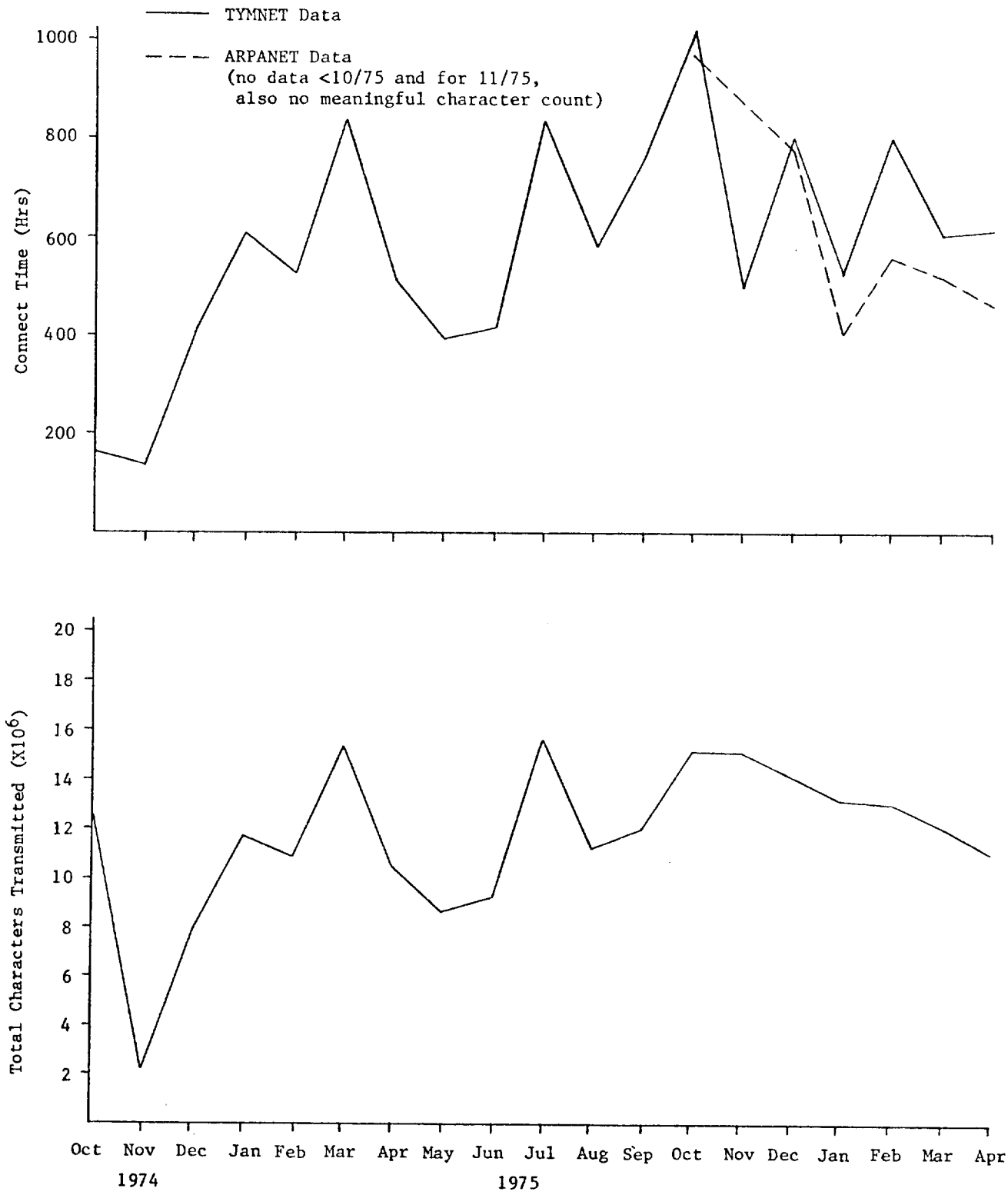
1)	Staff	50.99	2126	14453
2)	System & Operations	63.71	3688	27033
		---	---	---
	COMMUNITY TOTALS	114.70	5814	41486
		======	======	=====
	RESOURCE TOTALS	275.41	9732	93393

II.B.3 NETWORK USAGE STATISTICS

NETWORK USAGE PLOTS

The plots in Figure 10 show the major billing components for SUMEX-AIM TYMNET usage. These include the total connect time for terminals coming into SUMEX and the total number of characters transmitted over the net. The ratio of characters received at SUMEX to characters sent to the terminal is about 1:(10-14) over the past couple of months. Also shown for recent months is a plot of ARPANET connect time which tracks the corresponding data for TYMNET usage fairly closely. No data for "Character" transmission is available for ARPANET since file transfers and terminal traffic use different byte sizes and these data are not resolved and maintained for the ARPANET.

Figure 10.
SUMEX-AIM NETWORK USAGE



II.C RESOURCE EQUIPMENT SUMMARY

The following table gives a list of the items of equipment purchased to date for the SUMEX resource along with details on vendor, description, price, and date.

ITEM	QUANTITY	DESCRIPTION	MANUFACTURER	MODEL NUMBER	DATE INSTALLED	DATE ACCEPTED	PURCHASE PRICE	SOURCE FUNDS
KI-10 CPU	1	Central processor, including console	Digital Equipment Corporation	KI-10	3/1/74	4/24/74	$178,500	NIH
	1	Central processor, including console	Digital Equipment Corporation	KI-10	4/15/76	5/7/76	$203,138	NIH
Memory	3	Core memory (64K words including 4 MC-10 memory ports)	Digital Equipment Corporation	MF-10G	3/1/74	4/24/74	$224,910	NIH
	1	Core memory (64K words including 4 MC-10 memory ports)	Digital Equipment Corporation	MF-10G	11/74	12/74	$ 63,484	NIH
	1	Memory port multiplexer	Digital Equipment Corporation	MX-10	8/74	9/74	$ 4,770	NIH
Clock	1	Programmable clock	Digital Equipment Corporation	DK-10	3/1/74	4/24/74	$ 2,678	NIH
	1	Programmable clock	Digital Equipment Corporation	DK-10	4/15/76	5/7/76	(incl. in second processor)	
Disk System	1	Single double density disk controller	Digital Equipment Corporation	RP-10C	3/1/74	5/1/74		
	1	Memory data channel	Digital Equipment Corporation	DF-10	3/1/74	4/24/74		
	4	Double density disk drives and disk packs	Digital Equipment Corporation	RP-03	3/1/74	4/24/74	$108,153	NIH
	3	Double density disk drives and disk packs	Digital Equipment Corporation	RP-03R	2/75	3/75	$ 44,636	NIH

ITEM	QUANTITY	DESCRIPTION	MANUFACTURER	MODEL NUMBER	DATE INSTALLED	DATE ACCEPTED	PURCHASE PRICE	SOURCE FUNDS
Swapping Storage	2	Fixed head disk with 1.7M word capacity and 4 track parallel access	Digital Development Corporation	A-7312-D-8	1/75	3/75	$ 37,206	NIH
	1	Special systems controller for DDC disks	Digital Equipment Corporation	RES-10	10/74	11/74	$ 81,090	NIH
DEC Tapes (TU-56)	1	DEC tape control	Digital Equipment Corporation	TD-10	3/1/74	4/24/74	↓	↓
	1	Dual DEC tape drive	Digital Equipment Corporation	TU-56	3/1/74	4/24/74	$ 17,850	NIH
Magnetic Tapes (2 x TU-30)	1	Magnetic tape controller	Digital Equipment Corporation	TM-10A	3/1/74	4/24/74	↓	↓
	2	Tape transports	Digital Equipment Corporation	TU-30	3/1/74	4/24/74	$ 31,238	NIH
Line Printer	1	Special systems line printer control for Data Products 2410	Digital Equipment Corporation	Special	6/74	7/74	$ 7,208	NIH
	1	Line printer with 96 character drum, vertical format control, parity check	Data Products	2410	6/74	7/74	$ 18,963	NIH

ITEM	QUANTITY	DESCRIPTION	MANUFACTURER	MODEL NUMBER	DATE INSTALLED	DATE ACCEPTED	PURCHASE PRICE	SOURCE FUNDS
GT-40	1	Graphics terminal	Digital Equipment Corporation	GT-40	3/1/74	4/24/74	$ 11,156	NIH
Line Scanner	1	Data line scanner	Digital Equipment Corporation	DC-10A	3/1/74	4/24/74		
	1	8-line unit	Digital Equipment Corporation	DC-10B	3/1/74	4/24/74	$ 16,275	NIH
TYMNET Interface	1	PDP-10 TYMNET communications controller	TYMSHARE		8/74	10/74	$ 50,774	NIH
ARPANET Interface	1	BB&N ARPANET/KI-10 interface and VDH	Bolt, Beranek & Newman		1/75	2/75	$ 21,200	NIH
PDP-11/10	1	Communications processor	Digital Equipment Corporation	PDP-11/10	2/75	3/75	$ 13,445	NIH

ITEM	QUANTITY	DESCRIPTION	MANUFACTURER	MODEL NUMBER	DATE INSTALLED	DATE ACCEPTED	PURCHASE PRICE	SOURCE FUNDS
Terminals	1	Terminal	Data Terminals Communications	DTC-300	3/18/74	6/74	$ 4,597	NIH
	2	Terminals - Execuport portable with carry case	Computer Transceiver Systems, Inc.	311-3	3/18/74	6/74	$ 6,402	NIH
	6	Terminals - elite CRT with edit capabilities	Datamedia	2500	9-10/74	11/74		
	2	Terminals - elite CRT with edit capabilities	Datamedia	2500	8/75	8/75		
	2	Terminals - elite CRT with edit capabilities	Datamedia	2500	12/10/75	1/29/76	$ 24,618	NIH
Keyboards	3	Keyboards, special, for leased Datamedia elite 2500 CRT terminals at - NIH Rutgers Univ. Washington Univ.	Datamedia	70DVK7019	11/74	12/74	$ 1,118	NIH

ITEM	QUANTITY	DESCRIPTION	MANUFACTURER	MODEL NUMBER	DATE INSTALLED	DATE ACCEPTED	PURCHASE PRICE	SOURCE FUNDS
Modems	16	Auto answer modems	Prentice Electronics	P-113B	5/6/74	5/6/74		
	5	Auto answer modems	Prentice Electronics	P-1200/150	5/6/74	5/6/74		
	3	Auto answer modems	Prentice Electronics	P-1200/150	4/2/76	4/2/76		
	5	Originate modems	Prentice Electronics	P-1200/150	5/6/74	5/6/74		
	4	Modem enclosure with loopback switch and cables	Prentice Electronics	P-100	5/6/74	5/6/74		
	4	Modem enclosures for 8 modems with cables, power supply, digital loopback, line loopback, indicator lights	Prentice Electronics	P-850	5/6/74	5/6/74		
	3	Acoustic coupler modems	Prentice Electronics	DC-22	3/74	3/74		
	3	Modem enclosure with line loopback switch to house P-103F modems	Prentice Electronics	P-100	3/74	3/74	$ 11,319	NIH
Oscilloscope	1	Oscilloscope	Tektronix, Inc.	475DM43	1/75	1/75	$ 3,476	NIH

II.D PUBLICATIONS

Publications for the SUMEX staff have included papers describing the SUMEX-AIM resource and on-going research:

[1] Carhart, R.E., Johnson, S.M., Smith, D.H., Buchanan, B.G., Dromey, R.G., and Lederberg, J, Networking and a Collaborative Research Community: a Case Study Using the DENDRAL Programs, ACS Symposium Series, Number 19, COMPUTER NETWORKING AND CHEMISTRY, Peter Lykos (Editor), 1975.

[2] Levinthal, E.C., Carhart, R.E., Johnson, S.M., and Lederberg, J., When Computers Talk to Computers, Industrial Research, November 1975

Mr. Clark Wilcox was asked to chair the session on Languages for Portability at the DECUS DECsystem10 Spring '76 Symposium. A description of his work will appear in the proceedings.

In addition as reported earlier, a substantial effort has gone into developing, upgrading, and extending documentation about the SUMEX-AIM resource, the SUMEX-TENEX system, and the many subsystems available to users. These efforts include a number of major documents (such as SOS, PUB, and TENEX-SAIL manuals) as well as a much larger number of document upgrades, user information and introductory notes, an ARPANET Resource Handbook entry (see Appendix G), and policy guidelines (see Appendix F, Appendix I, and Appendix J). Publications for individual user projects are summarized in the respective reports (see Section IV).

III RESOURCE FINANCES

III.A REFERENCE TO BUDGETARY DETAILS

The budgetary materials for the SUMEX project covering past actual costs, current performance, and estimates for the next grant year are submitted in a separate document to the NIH.

III.B RESOURCE FUNDING

The SUMEX-AIM resource is essentially wholly funded by the Biotechnology Resources Program [*]. The various collaborator projects which use SUMEX are independently funded with respect to their manpower and operating expenses. They obtain from SUMEX, without charge, access to the computing and, in most cases, communications facilities in exchange for their participation in the scientific and community building goals of SUMEX.

[*] Except for the participation by Stanford University in accordance with general cost-sharing, and for assistance to SUMEX by other projects with overlapping aims and interests.

IV RESOURCE PROJECT DESCRIPTIONS

The following are inputs from the various user projects currently in the SUMEX-AIM community. These project descriptions and comments are the result of a solicitation for contributions sent to each of the project Principal Investigators requesting the following information:

I) Summary of research program

- A) Technical goals
- B) Medical relevance and collaboration
- C) Progress and accomplishments
- D) Current list of project publications
- E) Funding status (current funding level and pending applications or renewals)

II) Interactions with the SUMEX-AIM Resource

- A) Examples of collaborations and medical use of programs through networks
- B) Useful contacts and cross fertilization with other SUMEX-AIM projects (via workshop, messages, terminal links, etc.)
- C) Critique of resource services

The text which follows on the various projects is primarily the responsibility of the indicated project leaders.

IV.A FORMALLY APPROVED PROJECTS

IV.A.1 STANFORD USERS

IV.A.1.a DENDRAL PROJECT

DENDRAL PROJECT
Principal Investigators: Profs. C. Djerassi (Chemistry), J. Lederberg (Genetics), and E. Feigenbaum (Comp. Sci.)

(Grant NIH RR-00612-06, 3 years, $240,967 this year)

OVERVIEW

In the period August,1975 to July,1976 the DENDRAL programs and the gas chromatography/mass spectrometry (GC/MS) data system have made significant progress toward the goals stated in the research proposal. This report of progress is organized in three parts, corresponding to the three specific aims of our December, 1973, proposal: (PART 1) Enhancing the power of the mass spectrometry resource, (PART 2) Developing performance and theory formation programs, and (PART 3) Applying the computer programs and instrumentation to biomedically relevant structure elucidation problems.

The DENDRAL project, one of the major users at Stanford of the SUMEX-AIM computer facility, has also been forming its own community of remote users. This national "EXODENDRAL" community has already provided valuable contributions to program development and both the community and contributions are expected to grow at an increased rate.

PART 1: ENHANCING THE POWER OF THE M.S. RESOURCE

1.1 Introduction

Our grant proposal requested funds for significant upgrading of our capabilities in mass spectrometry. The goals of this upgrading were to provide routine high resolution mass spectrometry (HRMS), combined gas chromatography/low resolution mass spectrometry (GC/LRMS) and to develop a combined gas chromatography/high resolution mass spectrometry (GC/HRMS) facility. In addition, this would provide the capability for new experiments in the detection and utilization of data on metastable ions. These capabilities would then be available as required for application to our wider goal, solution of biomedical structure elucidation problems of a community of researchers.

The upgrading included several items of hardware and software development, as follows: 1) Acquire stand-alone computer support for the mass spectrometer because existing facilities were inadequate and very expensive; 2) convert existing software, written in the PL/ACME language

into FORTRAN so that it would run on the new system; 3) develop new software as required for the demanding task of GC/HRMS; 4) provide hardware and software for semi-automatic acquisition of data on metastable ions. The initial development phase of this upgrading included performance tests to determine the capabilities and limitations of the GC/HRMS system to define the scope of problems to which it can be applied. The past year's efforts (year two of the DENDRAL grant) have culminated in accomplishment of many of the above goals for development. In the first year, the computer system (a Digital Equipment Corp. PDP 11/45) was purchased, installed and is now operating routinely in conjunction with the mass spectrometer (a Varian-MAT 711) and an auxiliary PDP 11/20 system. Program conversion and modification for the initial version of the software system was completed and the computer system now provides complete stand-alone support for our experiments in mass spectrometry. Over the past year we have developed further our philosophy of data acquisition and reduction based on computed models of the actual performance of the mass spectrometer. This was and is necessary for routine automated collection and reduction of combined GC/HRMS data with minimal operator intervention in the procedures.

The system development is motivated by two goals. First, the system must be robust in the sense that it continue to operate under a variety of changing conditions, including intermittent misbehavior of the mass spectrometer. This ensures that the system can recover from hardware or software error conditions to prevent fatal "crashes" of the system and resulting loss of data. Second, the system must automate the GC/HRMS task. The volume of data acquired in GC/HRMS experiments can be efficiently handled only when every spectrum can be acquired and reduced for final output by the system without manual intervention. We are successful in these goals because we have written the software to determine the actual performance of the mass spectrometer and to have subsequent calculations based on that measured performance, as opposed to some hypothetical ideal.

We are now providing GC/HRMS service on a limited basis as we improve the system. The time devoted to system development and testing will slowly diminish over the next year, leaving additional time for analysis of mixtures obtained in our own work and that of our collaborators. We have deferred implementation of the metastable system (see below) while the GC/HRMS development is continuing, although we have completed the hardware and much of the software for the system.

Because we view GC/HRMS as the most important new capability of our mass spectrometer/computer work, the requirements of GC/HRMS have guided development of the software system. These requirements include continuous automatic monitoring of instrument performance to avoid wasting time collecting poor or erroneous data. By approaching GC/HRMS with an electrical recording system, we can monitor the instrument continuously, both during initial setup and during the course of the GC/HRMS experiment. While photographic recording may capture more of the signal, it is vulnerable to fluctuations in sample and instrument behavior in addition to the difficulties in reading the data from film for computer analysis. Major sections of the software and how they interact among one another are summarized below.

During the past year the routine production usage of the HRMS data has become a reality. The direct utilization of the system for the acquisition of high resolution mass spectrometry data typically occupies 6 hours per day. This figure does not include time for the post-processing of data, retrieval of data from the archival data base, or for the generation of duplicate print outs of selected data. These demands add 1 to 2 hours of system service each day to the total high resolution system requirements.

Low resolution mass spectral data whether it be smoothed from high resolution data or obtained directly as low resolution data, places additional time demands upon the data system. High to low resolution conversion, low resolution plotting, and low resolution spectral library searching have all generated a need for increasing amounts of system time.

In an effort to utilize the data system more completely during non-prime time, batch and spooling mechanisms have been constructed. The high resolution spectral reviewing mechanism may be actuated and then left unattended while the hard-copies are being generated. The high to low resolution conversion process contains a mechanism for the generation of a low resolution plotting spool which can be played without operator intervention. Batch procedures have been written which provide for the archival of newly acquired spectral data in the archival data base.

As with any system as large as the high resolution system there is a continual need for system maintenance and minor software upgrades. A wider range of data acquisition and analysis places new demands upon the system which require further modification of the software.

The net result of the production demands has been to reduce the amount of system time available for the development of new software facilities. Software development and production compete for the available system time reducing the productivity of both the chemical user and the software developer. This competition can be drastically reduced if software development can proceed on a machine separate from that on which production is done. The SUMEX PDP-10 and TENEX operating system provide a more tractable medium for development than does the restricted environment of the PDP-11.

A major factor in the ease with which programs can be constructed is the ease with which text can be manipulated. The TV-EDIT program which is available on the PDP-10 has proven to be effective for this task. This program provides an extremely flexible text editing system for display terminals. The mechanics of program construction can be greatly simplified by the utilization of this facility. Typically all major (more than a few changes) text modification of programs are carried out on the PDP-10 using TV-EDIT and then transferred to the PDP-11. Thus even the task of writing FORTRAN programs is simplified even though there exist FORTRAN incompatibilitier between the two machines.

While TV-EDIT has reduced development demands on the PDP-11 by eliminating PDP-11 text editing sessions, the problem of program compilation and debugging remain. Clark Wilcox, of the SUMEX staff, has provided an effective solution to this problem with the development of the

MAINSAIL (Machine Independent SAIL) compiler. This compiler provides the user with a powerful machine independent structured language. Not only is the compiler machine independent, but exhibits superior execution speeds and storage requirements as compared to the DOS 9 FORTRAN which has been used previously.

The combination of TV-EDIT and MAINSAIL has proven to be an effective method for the development of software for the PDP-11s within the PDP-10 environment. Most debugging can be carried out on the PDP-10 and then transferred to the PDP-11s for final debugging of machine-dependent facilities. The class of machine-dependent facilities includes device drivers and interaction with the operating system. The class of machine-independent facilities includes analysis algorithms, file manipulation, and most other programs which need development. This means that the amount of time required on the PDP-11 for program development can be reduced significantly using the aforementioned process, leaving more time for production demands.

1.2 Summary

As the above hardware and software improvements are being made we will continue evaluation of the GC/HRMS system in parallel with its actual application to real problems. GC/HRMS is a relatively new and difficult technique for routine application. In order to use it effectively, we will have to exert some effort toward determining and optimizing the performance of the many elements of the system, the GC, the MS, and the computer hardware and software.

PART 2: DEVELOPING PERFORMANCE AND THEORY FORMATION PROGRAMS TO ASSIST IN BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS

2.1 Introduction

The Heuristic DENDRAL computer programs assist with structure elucidation problems by helping interpret mass spectra and helping generate structures that are consistent with data obtained from a variety of spectroscopic and physical/chemical courses. The Meta-DENDRAL programs assist with rule formation problems in cases where the rules of mass spectrometry are not known.

Both the interpretation and rule formation programs are written as interactive tools to be controlled by professionals to combine the professional's judgment with the computer's combinatorial power.

2.2 CONGEN.

The CONGEN[48,53] program represents a significant extension of a program which has developed over the last several years, the cyclic structure generator[40,41]. The purpose of CONGEN is to assist the chemist in determining the chemical structure of an unknown compound by 1)

allowing him to specify certain types of structural information about the compound which he has determined from any source (e.g., spectroscopy, chemical degradation, method of isolation, etc.) and 2) generating an exhaustive and non-redundant list of structures that are consistent with the information. The generation is a stepwise process, and the program allows interaction at every stage; based upon partial results the chemist may be reminded of additional information which he can specify, thus limiting further the number of final structures.

CONGEN fits with the other DENDRAL programs as a "backstop" solution to structure elucidation problems. If the mass spectrum of an unknown compound is available, then CLEANUP and MOLION could be used, but if the general class of the compound is not known, PLANNER has no starting point from which to work. In such cases, structural information can be extracted manually from the spectrum and given to CONGEN for analysis. Because CONGEN makes no assumptions about the source of this information, other spectroscopic or chemical techniques may be used to supply supplemental data.

At the heart of CONGEN are two algorithms whose accuracy has been mathematically proven and whose computer implementation has been well tested. The structure generation algorithm[31,37,40,41] is designed to determine all topologically unique ways of assembling a given set of atoms, each with an associated valence, into molecular structures. The atoms may be chemical atoms with standard chemical valences, or they may be names representing molecular fragments ("superatoms") of any desired complexity, where the valence corresponds to the total number of bonding sites available within the superatom. Because the structure generation algorithm can produce only structures in which the superatoms appear as single atoms (we refer to these as intermediate structures), a second procedure, the imbedding algorithm[48,53] is needed to expand the superatoms to their full chemical identities.

These two routines give the chemist the ability to construct structures from a given set of molecular "building blocks" which may be atoms or larger fragments. By itself, this capacity is of limited utility because the number of final structures can be overwhelming in many cases. Usually, the chemist has additional information (if only some general rules about chemical stability, which the program has no concept of) that can be used to limit the number of structural possibilities. For example, he may know that because of a compound's stability, it cannot contain a peroxide linkage (O-O) and thus the programs need not consider such structures when there are two or more oxygens in the "building block" list.

In the past year CONGEN has reached the level of a practical production program which can aid chemists, both locally and at remote network sites, in solving the structures of drug-related compounds and natural products. The development of this program during the year has been strongly guided by the difficulties and new requirements which have appeared as it was applied to a wide variety of cases, and its efficiency and usefulness have increased dramatically. We report here the details of the modifications and additions we have made to CONGEN, and the effects they have had on its utility. Also, because of the rich repertoire of

structure modification and testing functions available within CONGEN, we have found it to be an invaluable "laboratory" for the testing of new ideas, and we briefly describe several pilot projects which form the basis for future research. Discussion of applications of CONGEN to problems of biochemical interest is included in Part 3.

NEW CAPABILITIES FOR THE USER. There have been several additions to CONGEN which are visible to the user and which generally increase the flexibility and power of the program. These include:

1) Making CONGEN aware of aromaticity, a chemical property of molecules which results from certain combinations of double bonds in rings. Aromaticity has a profound effect upon both the chemical reactivity and symmetry properties of molecules, and CONGEN can now be directed to detect aromaticity in its output structures, to compensate for the difference between the actual symmetry of an aromatic system and the symmetry which appears in the graph representing it, and to distinguish aromatic from non-aromatic atoms when it tests GOODLIST and BADLIST entries.

2) Giving the user the ability to type "?" to any prompt in the program, which results in a summary of the possible inputs. In some cases this summary is a list of possible commands, while in others it is a short explanatory message. A new interactive teletype-input routine was developed which makes it easy to include such help messages in the program, and which mimics the handy command-recognition and command-completion features of the TENEX operation system.

3) Including new specifications in the EDITSTRUC language for describing substructural features. The user can now declare a bond in a substructure to be an "anybond", which means that the atoms at the termini are connected but that the multiplicity of the connection is unspecified. This is especially handy when defining substructures containing aromatic portions because bond multiplicity is an indistinct concept in aromatic systems. Another new structural element which can be specified is a "linknode", a node which stands for a variable-length chain of atoms of the given type rather than a single atom. The minimum and maximum lengths of such a chain can be specified as well. The linknode feature is useful for defining constraints on ring fusions and other constraints such as Bredt's rule which depend on path length. Other extensions have been made internal to CONGEN which will shortly be reflected in the user-level language of EDITSTRUC. These include numerical inequalities involving node properties (e.g., "the number of H's on atom 3 is greater than the number of H's on atom 5") or linknode lengths (e.g., "the sum of the lengths of linknodes 2 and 6 is greater than 5"), and greater control over the number of fittings found for a GOODLIST constraint (e.g., the ability to distinguish between "the number of N's in six-membered rings" and "the number of six-membered rings containing N").

4) Allowing greater flexibility in the selection of terminal type. This choice controls the output of structural drawings so they are best suited to the user's terminal. Several different types of character-oriented and graphics-display terminals are now supported.

5) Making CONGEN accessible from the GUEST login account at SUMEX. This involved preventing a GUEST user from reaching certain critical points in CONGEN which would allow greater system access than is normally authorized for guests. We can now offer trial access to CONGEN via the guest mechanism without worrying about SUMEX misuse.

6) Creating a BATCH command for CONGEN. This allows the user to submit time-consuming, compute-bound calculations to the batch-processing facility of SUMEX. The computation is then run automatically at off-hours when it will not overload the system resources. The user can now run CONGEN in its interactive mode to input all of his data and then submit the large tasks to BATCH for overnite processing.

7) Including a pruning function MSPRUNE which is used to test a list of candidate structures for consistency with a set of observed peaks from a mass spectrum. The candidates are typically generated by CONGEN using structural data from other sources. The user specifies the observed MS peaks (high- or low-resolution, or a combination of both) along with a set of constraints on the allowed cleavage processes. MSPRUNE retains only those candidates which can account for the observations via one of these allowed processes. The constraints speak of the number of bonds broken and the number of steps in a process, the proximity of pairs of cleaved bonds (i.e., whether or not two adjacent bonds can break in a given process), the multiplicity or aromaticity of each cleaved bond and the possible neutral transfers. MSPRUNE is the first CONGEN function which can aid directly in the interpretation of "raw" spectral data.

2.3 Meta-dendral Rule Formation Programs

The INTSUM program [34] is in routine, production use to assist in interpretation of the mass spectra of new classes of molecules (see Part 3 for details). When the mass spectrometry rules for a given class of compounds are not known, the INTSUM, RULEGEN and RULEMOD programs can help a chemist formulate those rules. Essentially, these programs categorize the plausible fragmentations for a class of compounds by looking at the mass spectra of several molecules in the class. All molecules are assumed to belong to one class whose skeletal structure must be specified. Also, the mass spectra and the structures of all the molecules must be given to the program.

INTSUM collects evidence for all possible fragmentations (within user-specified constraints) and summarizes the results. For example, a user may be interested in all fragmentations involving one or two bonds, but not three; aromatic rings may be known to be unfragmented ; and the user may be interested only in fragmentations resulting in an ion containing a heteroatom. Under these constraints, the program correlates all peaks in the mass spectra with all possible fragmentations. The summary of results shows the number of molecules in whose spectra there is evidence for each particular fragmentation, along with the total (and average) ion current associated with the fragmentation.

The RULEGEN program attempts to explain the regularities found by