Table of Contents

List of Figures

List of Tables

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
National Institutes of Health

Division of Research Resources
Biomedical Research Technology Program
Annual Progress Report
PART I, TITLE PAGE

1. PHS AWARD NUMBER: 5 U 4 1 R R 0 1 6 8 5 - 0 3

2. TITLE OF AWARD: BIONET NATIONAL COMPUTER RESOURCE FOR MOLECULAR BIOLOGY

3. NAME OF RECIPIENT INSTITUTION: IntelliGenetics, Inc.

4. HEALTH PROFESSIONAL SCHOOL (If applicable):

5. REPORTING PERIOD:

5a. FROM (Month, Day, Year): 0 3 - 0 1 - 8 6

5b. TO (Month, Day, Year): 0 2 - 2 8 - 8 7

6. PRINCIPAL INVESTIGATOR:

6a. NAME: Dr. Michael J. Kelly

6b. TITLE: President, IntelliGenetics

6c. SIGNATURE: Michael J. Kelly

7. DATE SIGNED (Month, Day, Year): 12-12-86

8. TELEPHONE (Include Area Code): 4 1 5 - 9 6 5 - 5 5 9 0

Part II. Description of Program Activities

This section of our Annual Report provides statistical information on the use of the BIONET^{tm} Resource. The period covered is 12/85 - 11/86, to coincide with the dates of preparation of our Report and to follow our procedure of providing a full year's statistical information to compare with previous years' Reports.

Individual sections are prepared under guidelines discussed previously with BRTP staff and used in our previous Reports. We use a format for reporting the hundreds of individual Principal Investigators' use that is easy for us to generate while retaining the critical information necessary for BRTP in its internal and governmental reporting requirements. Complete research abstracts are kept at IntelliGenetics and are available upon request.

The BIONET User community is divided into different classes, representing different levels of use of the computer system and staff resources, as follows:

- **Class I.** Class I users represent the Service component of the scientific community. They participate in the electronic communications facilities of BIONET (bulletin boards and electronic mail), and use the Core and Contributed Software libraries to pursue their research;

- **Class II.** Class II users represent the Collaborative component of the user community. Scientists in Class II enjoy all benefits of Class I use, and in addition contribute software and expertise to BIONET, working closely with BIONET staff. This category also includes bulletin board leaders, accounts by courtesy with other, related Resources (GenBank, NBRF/PIR, Dana Farber, etc.), National Advisory Committee members and accounts for communication with BIONET Satellites.

- **Class III.** The category of Class III access has been reserved for system managers of local computer facilities. Such persons might not qualify as Principal Investigators, but are willing to work closely with other researchers of BIONET at a local site to help them learn to use the system and telecommunications effectively. They share Class II privileges.

- **Class IV.** Class IV users consist of those scientists who wish access only to the electronic communication facilities of BIONET. They are given access to the electronic mail and bulletin board facilities.

Information on number of PI's by Class is summarized in Table II-1.

The total number of investigators with access to BIONET, 489, is about 40 less than the total presented in our last annual report. During the past year about 170 investigators chose not to renew their accounts primarily because of their anticipated lack of need for access to BIONET and the imposition of the subscription fee. During the year we have added about 130 new investigators to the Resource. We expect that the trend of bringing substantial numbers of new users on to the system will continue. Attrition in the future will not, however, be as heavy because all new users come to BIONET with full knowledge of

Table II-1: Summary of the BIONET User Community

Class I	444
Class II	37
Class III	4
Class IV	4
Total	489

the subscription fee. Rather, we expect that the growing use of locally-available sequence analysis software, provided through the BIONET Satellite program or PI's purchase of microcomputer software, will diminish the load per PI on BIONET, enabling us to support a larger number of users.

II.A. Scientific Subprojects

II.A.1. Collaborative Research and Service

In the following section we report the use of the BIONET Resource for Class I-IV users. The "Usage Factor" is reported as both central processor unit (cpu) time, in minutes and connect time in hours, for each Principal Investigator. These values are the sum of all usage by the PI and his or her group members ("Sub-I's"). We report data only on those PI groups that have used the Resource during the past 12 months. Of the 489 PI's, 418, representing about 1463 individual investigators, have accessed BIONET. Detailed statistics on the use by each individual are maintained by the BIONET computer and are available to interested parties.

We do not report Resource staff hours nor BRTP funds allocated for individual PI's because it is impossible to allocate these rationally to such a large user community. Summary information on allocation of staff hours is given in Section II.C, the Resource Summary Table.

II.A.2. Core Research and Development

We report on the standard form the summary information for our Core Research projects.

The Resource Technology used is the DEC-2060 computer for all projects.

The Usage Factor is reported as minutes of cpu time used for the project. This number is derived for each investigator by multiplying the fractional time ("FT") spent on the project times the total cpu time used during the past twelve months.

Resource Staff Hours are based on the same FT multiplied times the total hours spent on BIONET for the last twelve months.

"BRTP Funds Allocated" are calculated as followed, from the sum of the following components:

- Actual Personnel Costs. The personnel costs for each project are derived by multiplying the above FT for each BIONET staff person's time spent on the project times their respective annual salary plus fringe benefits; the actual personnel cost is the sum of these individual figures.

- Consultant Costs. The FT spent by a Co-Investigator involved in a project is multiplied times the total consulting cost for the Co-I; these are summed for each project where appropriate.

- Fraction of Awarded Funds. The fraction of total awarded funds for each project is derived by multiplying the fractional time spent on the project by the awarded funds (defined below). The fractional time is determined from the sum of hours spent on the project by all investigators divided by the sum of hours spent on BIONET by all investigators. In computation of awarded funds, we include the grant categories of *Supplies*, *Travel*, and *Other Expenses*. The categories of *Personnel* and *Consultants* are accounted for in the previous two computations. For this calculation we have used the actual time spent in the last twelve months and a cost basis of the estimated total expenditures in the above categories for this grant year. Although these are three months out of phase, we do not think the fractional time spent will change significantly during the next three months of the current grant period.

DRR SCIENTIFIC SUBPROJECT FORM

PART II, SECTION A — **INSTITUTION:** IntelliGenetics

AWARD NUMBER 5 U 4 1 R R 0 1 6 8 5 — 0 3 — **REPORT PERIOD:** March 1, 1986 to February 28, 1987

Fill out a separate Subproject Form for Core, Collaborative or Training. Check one of the following:

[X] CORE RESEARCH & DEVELOPMENT [ ] COLLABORATIVE RESEARCH & SERVICE [ ] TRAINING

(1) Descriptive Title (80 characters) Abstract	(2) Science Code Axis I	(2) Science Code Axis II	(3) a. Investigator(s) Name (Last Name, First Name & Middle Initial) b. Department c. Non-Host Inst.	(4) USAGE FACTOR Resource Technology a/	CPU MIN USED	Resource Staff Hours	(5) BRTP Funds Allocated
Multiple Sequence Alignment Collection, Porting, and Evaluation of several computer programs for intercomparison of multiple (3 or more) biological sequences; Release of validated programs to the BIONET community with a review of strengths and weaknesses.	9	42, 68	a. Smith, Dennis H. Brutlag, Douglas L. Friedmann, Theresa A. b. IntelliGenetics	DEC-2060 " "	12 13 80	55 10 56	8526
BIONET Satellite Communications Districation of the computational and communication facilities of BIONET to other, geographically remote sites (Satellite Resources); linking these Satellites for rapid exchange of Electronic Mail and Bulletins, Data Files and Programs.	9	40, 42	a. Smith, Dennis H. Roode, David R. Relph, John R. Liebschutz, Robert Friedland, Peter Boyd, Brian Levy, Benjy b. IntelliGenetics	DEC-2060 " " " " " "	12 156 1289 137 3 1 1	55 180 257 196 10 27 39	48,434
Special Hardware for Text Search Exploring new parallel processing technologies for high speed text search and sequence homologies for applications to biological sequence retrieval and intercomparison.	9	42, 68, 70	a. Smith, Dennis H. Friedland, Peter Brutlag, Douglas L. b. IntelliGenetics	DEC-2060 " "	35 33 40	166 96 29	26,425
CUMULATIVE TOTALS: No. subprojects	3				1810	1176	83,385

a/ Identify Resource Technologies Used. b/ Give Hours Resource Technologies Used. See Instructions, page ii.

II.A.3. Training

We report summary information for our Training program. The sites at which BIONET provided some level of training are named here and are discussed in more detail in Chapter III, *Narrative Description*, section III.A.4

The method for calculation of Usage Factors and BRTP Funds Allocated is the same as that described above under Core Research and Development.

DRR SCIENTIFIC SUBPROJECT FORM

PART II, SECTION A

INSTITUTION: IntelliGenetics

AWARD NUMBER: 5 U 4 1 R R 0 1 6 8 5 — 0 3

REPORT PERIOD: March 1, 1986 to February 28, 1987

Fill out a separate Subproject Form for Core, Collaborative or Training. Check one of the following:

[ ] CORE RESEARCH & DEVELOPMENT [ ] COLLABORATIVE RESEARCH & SERVICE [X] TRAINING

(1) Descriptive Title (80 characters) Abstract	(2) Science Code Axis I	Axis II	(3) a. Investigator(s) Name (Last Name, First Name & Middle Initial) b. Department c. Non-Host Inst.	(4) USAGE FACTOR Resource Technology a/	CPU MIN USED	Resource Staff Hours	(5) BRTP Funds Allocated
BIONET Training Program Support of traing for BIONET scientists, including phone trainings, preparation of new documentation for training, and outside trainings at Stanford University, Miami Mid-Winter Symposia, American Scoiety of Biological Chemists, University of New Hampshire, and the Macromolecules, Genes and Computers meeting.	9	40.68	a. Allen, Marcia	DEC-2060	24	72	
			Bigham, Nancy	"	162	115	
			Brutlag, Douglas L.	"	66	48	
			Friedmann, Theresa A.	"	172	119	
			Friedland, Peter	"	3	10	
			Kristofferson, David	"	41	24	
			Lawler, Maryjo	"	164	221	
			Smith, Dennis H.	"	12	55	41,012
CUMULATIVE TOTALS: No. subprojects	1				644	664	41,012

a/ Identify Resource Technologies Used. b/ Give Hours Resource Technologies Used. See Instructions, page 11.

II.B. Books, Papers, Abstracts

We report the publications by members of the BIONET scientific community on a version of the special form provided by BRTP. These publications have **ALL** arisen from use of BIONET, although support by BIONET and the NIH has not always been acknowledged.

The figures on *Cumulative Number Published* refer to the current year alone. This year we have received 114 publications resulting from the use of BIONET, a substantial increase over the 43 reports received last year.

Part II Section B Award Number: 5U41RR01685-03
INSTITUTION: IntelliGenetics Report Period: March 1, 1986 to February 28, 1986

COLLABORATIVE RESEARCH AND SERVICE

Cooke, N. E., David E., Vitamine D. Binding Protein is a Third Member of the Albumin and Alpha-Fetoprotein Gene Family. J. Clin. Invest. 76:2420-2424, 1985.

Cooke, N.E. Rat Vitamin D Binding Protein: Determination of the Full-Length Primary Structure from Cloned cDNA. J. Biol. Chem. 261:3441-3450, 1986.

Cooke, N. E., et al, Direct Regional Assignment of the Gene for Vitamin D. Binding Protein (Cc-Globulin) to Human Chromosome 4q11-4q13 and Identification of an Associated DNA Polymorphism. Human Genetics 73:335-229, 1986

Leibhaber, S.A., Ray, J., Cooke, N. E., Sythesis of Growth Hormore - Prolactin Chimeric Porteins and Processing Mutants by the Exchange and Deletion of Genomic Exons. J. Biol. Chem. 261:14, 301-14,306, 1986.

Elsholtz, Rosenfeld, et al, Two Different Cis-Active Elements Transfer the Transcription Effects of Both EGF and Phorbol Esters. Science In press.

Howe, C., Overton, C., et al, Expression of the Testis and Ovary is Regulated During Embryogenesis. In preparation.

Miller, R. H., Robinson, W., Common Evolutionary Origin of Hepatitis B Virus and Retroviruses. PNAS 83:2531-2535 1986

Keen, N., Tamaki, S., Structure of Two Pectate'yase Genes from Erwinia Chrysanthesis EC16 and Thier High Level Expression in Esckerichier Coli. J. Bacteriol. In press

Trumbly, R. J., Isolation of the Yeast CyC8 Gene Responsible for Catabolite Repression. In preparation.

Trumbly, R. J. Sequence Analysis of the Yeast CYC8 Gene Predicts a Highly Glutamine Rich Protein. In preparation.

Chen, L-S, Mullin, D., Newton A., Identification, Nucleotide Sequence and Control of Developmentally Regulated Promotors in the Hook Operon Region of Caulobacter crescentus. PNAS USA 83:3860-2864 1986.

Minnich, S., Newton, A., Promoter Mapping and Cell Cycle Regulation of Flagellin Gene Transcrpits in Caulobacter crescentus. PNAS USA In press.

Grimes, S. R. Jr., Daum H.A. III, Smith, J. M., et al A Histone H4 Gene is Closely Associated with the Testis-Specific H1t Gene. Nucl. Acids Resear. In preparation.

Strobel, M., Abelson, J. Intron Mutations Affect Splicing of S. Cervisiae SUP53 Precursor tRNA.
Molec. & Cell. Biol. 6:2674-2683 1986.

Part II Section B Award Number: 5U41RR01685-03
INSTITUTION: IntelliGenetics Report Period: March 1, 1986 to February 28, 1986

COLLABORATIVE RESEARCH AND SERVICE

Tschunko, A.H., Allen, S. L., Long Inverted and Direct Repeats Occur in the Vicinity of Eliminated Sequences of Tetrahymena thermophila. Nucl. Acids Resear. Submitted 1987.

Tschunko, A. H., Allen, S. L., Genomic Organization and Developmental Fate of Adjacent Repeated Sequences of Tetrahymena thermophila. Genetics Accepted 1987.

Lohe, A.R., Brutlag, D. L., Multiplicity of Satellite DNA Sequences in Drosophile Melanogaster. PNAS 83:696-700 1986.

Lohe, A. R., Brutlag, D. L., Adjacent Satellite DNA Segments in Drosophila : Structure of Functions. J. Mol. Biol. Accepted 1987.

Lohe, A. R., Brutlag, D. L., Identical Satellite DNA Sequences in Sibling Species of Drosophila. J. Mol. Biol. In press 1987

Pritchard, A.E., Seilhamer, J., Cummings, D., Paramecium mitochondrial DNA Sequences and RNA Transcripts for Cytochrome Oxidase Subunit I, RF1, and Three ORFs Adjacent to the Replication Origin. Gene 44:243-253 1986.

Mahelingam, R., Cummings, D., et al Identification of Paramecium mitochondrialproteins using antibodies Raised Against Fused Mitochondrial Gene Products. Gene In press 1986.

Davis, R.J., Johnson, D.J., Czech, M.P., et. al. Identification of Serine 24 as the Unique Site on the Transferrin Receptor Phosphorylated by Protein Kinase C. J. Biol. Chem. 261 9034-9041 1986

Krawetz, S.A., Connor, Wayne, Cannon, Paul D., and Gordon H. Dixon A Vector-Primer-Cloner-Sequencer Plasmid for the Construction of cDNA Libraries: Evidence for a Rat Glyceraldehyde-3-Phosphate Dehydrogenase-Like mRNA and a Ferritin mRNA within Testis. DNA 5:427-435.

Lovett, M., cheng, Z.Y., Epstein, C.J. et al Molecular Markers for the Agouti Coat Color Locus of the Mouse. Genetics In press.

Lincoln, J.E. Cordes, S., Read, E., and Fischer, R.L. Regulation of Gene Expression by Ethylene during Fruit Development. PNAS In press

Glimcher, L.H. and Griffith, I.J. Mutation of Class II MHC molecules. Immunology Today In press

Griffith, I.J., Choi, E., and Glimcher, L.H. A single mutation in a Class II A ∂ gene alters T cell recognition. Proc. Nat'l Acad. Sci. In press

Griffith, I.J., Carland, F.M., and Glimcher, L.H. Alteration of a non-polymorphic residue in a Class II Eß gene eliminates an

COLLABORATIVE RESEARCH AND SERVICE

antibody-defined epitope without affecting T cell recognition. Submitted.

Biggs, J., Searles, L.L., and Greenleaf, A.L. Structure of the Eukaryotic Transcription Apparatus: Features of the Gene for the Largest Subunit of Drosophila RNA Polymerase II. Cell 42:611-621. 1985

Hamori, E. Novel DNA Sequence Representations. Nature 314: 585-586 1985

Hamori, Eugene and Zhou, Bing Analysis and Comparison of Two Fully Sequenced Viral DNA Genomes Using H Curves. Biophysical Journal 49, 125a 1986

Hamori, Eugene Analysis of Long DNA Sequences Using the Novel Method of H Curves (Abstr.) Hungarians in Science p. 768 Budapest, 1986 (MTA/MTESZ)

Hamori, Eugene Long Range Nucleotide-compostion Patterns in Fully-sequenced Genomes. Biol. Chem. Hoppe-Seyler 367S, 226 1986

Herrick, G., Cartinhour, S., Dawson, D. et al Mobile Elements bounded by C4A4 telomeric repeats in Oxytricha fallax. Cell 43:759-768 1985

Herrick, G., Cartinhour, S., Williams, K.R., and Kotter, K.P. Multiple Sequence Versions of the Oxytricha fallax 81-MAC alternate processing family. J. Protozool. In press.

Bray, S.J. & Hirsh, J. The D.virilis dopa decarboxylase gene is developmentally regulated when integrated in D. melanogaster. EMBO J. 5:2305-2312 1986

Scholnick, S., Bray, S.J., Hirsh, J., et al Central nervous system and hypoderm regulatory elements of the D. melanogaster dopa decarboxylase gene. Science In press

Morgan, D., Johnson, W.A. & Hirsh, J. Regulated splicing produces different forms of dopa decarboxylase in the central nervous system and hypoderm of Drosophila melanogaster. EMBO J. In press

Wiginton, D.A., Kaplan, D.J., Hutton, J.J. et al Complete Sequence and Structure of the Gene for Human Adenosine Deaminase. Biochemistry In press

Wabiko, H., Raymond, K.C., Bulla, L.A. Bacillus thuringiensis Entomocidal Protoxin Gene Sequence and Gene Product Analysis. DAN 5:305-314 1986

Aird, S.D., Kaiser, I.I., Lewis, R.V., Kruegel, W. Complete Amino Acid Sequence for the Basic Subunit of Crotoxin. Arch. Bich. and Biophys 249:296-30 1986

Part II Section B Award Number: 5U41RR01685-03
INSTITUTION: IntelliGenetics Report Period: March 1, 1986 to February 28, 1986

COLLABORATIVE RESEARCH AND SERVICE

Aird, S.D., Kaiser, I.I., Lewis, R.V. and Kruegel, W. Rattlesnake Presynaptic Neurotoxins: Primary Structure and Evolutionary Origin of the Acidic Subunit. Biochem. 24:7054-7058 1985

Chung-Her Jenh, Tiliang Deng, Johnson, Lee F., et al Mouse Thymidylate Synthase. Locations of Introns and Multiple Transciptional Start Sites. J. Biol. Chem In press 1986

Donohue, T. J., Hoger, J., and Kaplan, S., Cloning and Expression of the Rhodobacter sphaeroides Reaction Center H. Gene. J. Bact. 168: In press 1986

Donohue, T. J., McEwan, A., Kaplan, S. Cloning, DNA Sequence and Expression of the Rhodobacter sphaeroides Cytochrome c2 Gene. J. Bact. 168: In press 1986

Kemper, B., Molecular Biology of Parathyroid Hormone. Critical Rev. in Biochem. 19:353-380 1986

Govind, S., Bell, P.A., Kemper B., Structure of Genes in the cytochrome P-450PBc subfamily: Conservation of Intron Locations in the Phenobarbital-inducible Family. DNA 5: In press 1986

Hussain, M., Pastor, F., Lampen J.O., Cloning and Sequencing of blaz Encoding B-LactamaseIII, a Lipoprotein of Bacillus cereus 569/H. J. Bact. In press 1987

Boyer, P.D., Lengyel, J.A., Molecular Characterization of bsg25D: A Blastoderm-specific Locus in Drosophila melanogaster encoding a Product with Domains of Structural Similarity to Fos and Tropomyosin Proteins. Submitted

Wernke, S.M., Lingrel, J. B., Nucleotide Sequence of the Goat Embryonic ∂ Globin Gene (zeta) and Linkage and Evolutionary Analysis of the Complete ∂ Globin Cluster. J. Mol. Biol. 192: In press 1986

Hind K.R., Litman G. W., Major Reorganization of Immunoglobulin VH Segmental Elements During Vertebrate Evolution. Nature 320:546-549 1986

Kimmel, B. E., Yarbrough, L. R., et al, Tubulin Genes of the African Trypanosome Trypanosoma brucei rhodesiense: Nucleotide Sequence of a 3.7 kb Fragment Containing Genes for Alpha and Beta Tubulin. Gene 35:237-248 1985

Hirschberg, R., Yarbrough., L. R., et al, Cloning and Characterization of the Nitrogenase Genes from Anabaena veriabilis. J. Biotech. 2:23-38 1985

Wu, J., Yarbrough, L. R., et al, Tubulin Genes of the African Trypanosome: Nucleotide Sequence and Conservation of 5' Untranslated Regions of Tandem Repeats. Submitted

Part II Section B Award Number: 5U41RR01685-03
INSTITUTION: IntelliGenetics Report Period: March 1, 1986 to February 28, 1986

COLLABORATIVE RESEARCH AND SERVICE

Marsh, J., Eveleth, Evidence for Evolutionary Duplication of Genes in the Dopa Decarboxylase Region of Drosophila. Genetics 114:469-483 1986

Bland, M.M., Levings, C. S. III, and Matzinger, D. F., The Tobacco Mitochondrial ATPase Subunit 9 Gene is Closely Linked to an Open Reading Frame for a Ribosomal Protein. Mol. Gen. Genet. 204:8-16 1986

Miller, V. L., Mekalanos, J. J., Genetic Analysis of the Cholera Toxin-Positive Regulatory Gene toxR. J. Bact. 163:580-585 1985

Betley, M, Miller, V., Mekalanos, J. J., Genetics of Bacterial Exterotoxins. Ann. Rev. Microbio. 40:577-605 1986

Osawa, S., Miller, T. B. Jr., et al, Isolation of Partial cDNA's for Rat Liver and Muscle Glycogen Phosphorylase Isozymes. FEBS 202:282-288 1986

Oshima, R. G., Millan, J., Cecna, G., Comparison of Mouse and Human Keratin 18: A Component of Intermediate Filaments Expressed Prior to Implantation. Differentiation. In press

Bikenmeyer, L., Sugisaki, H., Ray, D. S., The Majority of Minicircle DNA in Crithidia fasciculata Strain CF-C1 is of a Single Class with Nearly Homogeneous DNA Sequences. Nucl. Acids Res. 13:7107-7118 1985

Birkenmeyer, L., Ray, D. S., Replication of kDNA in Kinetoplasts Isolated from Crithidia fasciculata: Identification of Minicircle DNA Replication Intermediates. J. Biol. Chem. 261:2362-2368 1986

Krusius, T., Ruoslahti, E., Primary Structure of an Extracellular Matrix Proteoglycan Core Protein Deduced From Cloned cDNA. PNAS USA 83:7683-7687 1986

Argraves, W., Pytela, R., Ruoslahti, E., et al, cDNA Sequences from the Alpha Subunit of the Fibronectin Receptor Predict a Transmembrane Domain and a Short Cytoplasmic Peptide. J. Biol. Chem. 261:12922-12924 1986

Suzuki, S., Ruoslahti, E., et al, cDNA and Amino Acid Sequences of the Cell Adhesive Protein Receptor Recognizing Vitronectin Reveal a Transmembrane Domain and Homologies with Other Adhesion Protein Receptors. PNAS USA In press

Bourdon, M., Shiga, M., Ruoslahti, E., Identification from cDNA of the Precursor Form of a Chondroitin Sulfate Proteoglycan Core Protein. J. Biol. Chem. 261:12534-12537 1986

Bourdon, M. , Shiga, M., Ruoslahti, E. Gene Expression of the Chondroitin Sulfate Proteoglycan Core Protein PG19. Mole. Cell. Biol. In press

Part II Section B Award Number: 5U41RR01685-03
INSTITUTION: IntelliGenetics Report Period: March 1, 1986 to February 28, 1986

COLLABORATIVE RESEARCH AND SERVICE

Poncz, M., Schwartz E., et al, Cloning and Characterization of Platelet Factor 4 cDNA Derived from a Human Erythroleukemic Cell Line. Blood 69: In press 1987

Sykes, P., Sokatch, J. R., et al, Cloning and Expression of Branched Chain Keto Acid Degydrogenase in Pseudomonas Putida and Escherichia Coli. Submitted 1986

Bova, C., Manfredi, J., Swanstrom R., env Genes of Avian Retroviruses: Nucleotide Sequence and Molecular Recombinants Define Host Range Determinants. Virology 152:"343-354 1986

Lerner, C., Switzer, R. L., Cloning and Structure of the Bacillus Subtilis Aspartate Transcarbamylase Gene (pyrB). J Biol. Chem. 261:11156-11165 1986

Theil, E. C., Ferritin: Structure, Gene Regulation and Cellular Function in Animals, Plants and Microorganisms. Ann. Rev. Biochem. 56: In press 1987

Roth, W., Wahba, A., et al, The Expression of a Gene for Mouse Eukaryotic Elongation Factor Tu During Murine Erythroleukemic Cell Differentiation. In preparation

Faber, S., Wilcox K., Association of the Herpes Simplex Virus Regulatory Protein IÇ∏4 with Specific Nucleotide Sequences in DNA. Nuc. Acids Res. 14:6067-6083 1986

Kuroda, M., Yanofsky, C., et al, Regulatory Elements Common tothe Bacillus Pumilus and Bacillus Subtilis trp Operons. J. Bact. 167:792-798 1986

Stanton, L., Schwab, M., Bishop, J.M., Nucleotide Sequence of Human N-myc. PNAS USA 83:1772-1776 1986

Gerondakis, S., Bishop, J.M., Structure of the Protein Encoded by the Chicken Proto-Oncogene c-myb. Mol. Cell. Biol. In press 1986

Ness, D. B., Grumet F.C., New Polymorphisms of HLA-B27 and Other B Locus Antigens Detected by RFLP Using a Locus Specific Probe. Human Immun. In press 1986

Wells, D., Bains, W., Kedes, L., Condon Usage in Histone Gene Families of Higher Eukaryotes Reflects Functional Rather than Phylogenetic Relationships. J. Mol. Evol. 23:224-241 1986

Erba, H., Gunning, P., Kedes, L., Nucleotide Sequences of the Human Upsilon Cytoskeletal Actin mRNA: Anomalous Evolution of Vertebrate Non-muscle Actin Genes. Nucl. Acid Res. 14:4275-5294 1986

Wade, R., Kedes, L., et al, Nucleotide Sequence, Tissue Specific Expression, and Chromosome Location of Human Carbonic Anhydrase

Part II Section B Award Number: 5U41RR01685-03
INSTITUTION: IntelliGenetics Report Period: March 1, 1986 to February 28, 1986

COLLABORATIVE RESEARCH AND SERVICE

III: The Human CAIII Gene is Located on the Same Chromosome as the Closely Linked CAI and CAII Genes. PNAS USA In press 1986

Regier, J. C., Evolution and Higher-order Structure of Architectural Proteins in Silkmoth Chorion. EMBO J. 5:1981-1989 1986

Duffy, L., Kurosky, A., Lai, C-Y., Cholera Toxin A Subunit: Functional Sites Correlated with Regions of Secondar Structure. Arch. Bio. Biophys. 235:549-555 1985

Nagel, G., Kurosky, A., et al, Evidence for the Expression of Three Genes Encoding Homologous Atrial Gland Peptides that Cause Egg Laying in Aplysia. J. Biol. Chem. 261:7853-7859 1986

Singh, S., Kurosky, A., Awasthi, Y., Human Liver Glutathoine S-transferase: Chemical Characterization and Secondary Structure Comparison with Other mamalian Glutathione S-transferases. Biochem. J. Submitted 1986

Yamanaka, M., Tobin, S., et al, Structure and Expression of the Drosophila Calmodulin Gene. Submitted 1986

Vigoreaux, J., Tobin, S., The 5C actin Gene of Drosophila melanogaster Encodes Multiple Transcrpits Sharing the Entire Protein Coding Sequence but Differing in their 5' and 3' Non-translated Sequences. In preparation 1987

Flink, I., Morkin, E., et al, Complete Amino Acid Sequence of Human Thyroxine Binding Globulin Deduced from Cloned DNA: Close Homology to the Serine Antiproteases. PNAS USA 83:7708-7712 1986

Bishop, D., Calhoun, E., et al, Human a-galactosidase A: Nucleotide Sequence of a cDNA Clone Encoding the Mature Enzyme. PNAS USA 83:4859-4863 1986

Lindahl, L., Zengel, J., Ribosonel Genes in Escherichia coli. Ann. Rev. Genet. 20:297-326 1986

Edenberg, H., Zhang, K., Isolation of a Mouse-Liver Alcohol Dehydrogenase cDNA. ICSU Short Reports 1:162-163 1984

Edenberg, H., Zhang, K., et al, Cloning and Sequencing of cDNA Encoding the Complete Mouse-Liver Alcohol Dehyrogenase. PNAS USA 82:2262-2266 1985

Crabb, D., Edenberg, H., Cloning and Sequencing of a cDNA Encoding Rat Liver Alcohol Dehydrogenase. Alcohol and Alcoholism In press 1986

Crabb, D., Edenberg, H., Complete Amino Acid Sequence of Rat Liver Alcohol Dehydrogenase Deduced from the cDNA Sequence. Gene In press 1986

Part II Section B Award Number: 5U41RR01685-03
INSTITUTION: IntelliGenetics Report Period: March 1, 1986 to February 28, 1986

COLLABORATIVE RESEARCH AND SERVICE

Leung, W., Liu, A., et al, Protein Engineering of Human Interferon Gamma by Using Novel Approaches for Chemical-Enzymatic Gene Synthesis and Fragment Replacement Method for Sit Specific Mutagenesis. Biol. of Interferon Sys. p. 412-424 1985

Liao, S., Leugn, W., Kanamaru, T., Structure and Epitope Specificity of Human Melanoma-Associated Oncofetal Antigen gP87 Analyzed with Monoclonal Antibodies. J. Dermat. 12:117-128 1985

Hasnain, S., Manavathu, E., Leung, W., DNA-mediated Transformation of Chlamydomonas Reinhardi Cells: Use of Amino-glycoside 3'-phosphotransferase as a Selectable Marker. Mol. and Cell. Biol. 5:3647-3650 1985

Hasnain, S., Leugn, W., et al, Development of Host Cell System for Production of Proteins of Biological and Medical Importance: Implications for Biotechnology. "Proteins in Food, Health and Industry" In press 1986

Jing, G., Liu, A., Leung, W., A Method for the Preparation of Size Marker for Synthetic Oligonucleotides. Analy. Biochem. 155:367-378 1986

Diaz-Mitoma, F., Leung, W., et al, Detection of Epstein Barr Virus in Throat Washings by DNA-DNA Dot Hybridization. J. Infec. Dis. In press 1986

Laksman, E., Hiruki, C., Wu, X., Leung, W., Use of [32p]-RNA Probes for the Dot-hybridization Detection of Potato Spindle Tubor Viroid. J. Virol. Meth. In press 1986

Hasnain, S., Manavathu, E., Leung, W., Construction of a Plasmid Vector Bearing the Initiation ATG Condon for Direct Expression in E. coli. Plasmid In press 1986

Leung, W., Manavathu, E., et al, Development of Fungal and Algal Cells for Expression of Foreign Genes. Applied Virology 2: Academic Press 1986

Kolbe, H., Wohlrab, H., Sequence of the N-terminal Formic Acid Fragment and Location of the N-Ethylmaleimide-binding Site of the Phosphate Transport Protein from Beef Heart Mitochondria. J. Biol. Chem. 260:15899-15906 1985

Rasmussen, V., Wohlrab, H., Bovine Cardiac Mitochondrial ADP/ATP-Carrier: Two Distinct mRNAs and an Unusually Short 3'-Noncoding Sequence. Biochem. Biophys. Res. Comm. 138:850-857 1986

Upton, C., McFadden G., DNA Sequence Homology between the Terminal Inverted Repeats of Shope Fibroma Virus and an Endogenous Cellular Plasmid Species. Mol. Cell. Biol. 6:265-276 1986

Upton, C., McFadden G., Tumorigenic Poxviruses: Analysis of Viral DNA Sequences Implicated in the Tumorigenicity of Shope Fibroma Virus and Malignant Rabbit Virus. Virology 152:308-321 1986

Part II Section B Award Number: 5U41RR01685-03
INSTITUTION: IntelliGenetics Report Period: March 1, 1986 to February 28, 1986

COLLABORATIVE RESEARCH AND SERVICE

Upton, C., McFadden, G., Identification and Nucleotide Sequence of the Thymideine Kinase Gene of Shope Fibroma Virus. J. Virol. In press 1986

Chang, W., McFadden, G., et al, Genome of Shope Fibroma Virus, a Tumorigenic Poxvirus, Contains a Growth Factor Gene with Sequence Similarity to those Encoding Epidermal Growth Factor and Transforming Growth Factor Alpha. Mol. Cell. Biol. In press

Upton, C., Carrell, R., McFadden, G., A Novel Member of the Serpin Superfamily is Encoded on a Circular Plasmid-like DNA Species Isolated from Rabbit Cells. FEBS Lett. In press

CUMMULATIVE # SUBMITTED		BOOKS 0	PAPERS	17	ABSTR.	0
CUMMULATIVE # PUBLISHED		BOOKS 0	PAPERS	67	ABSTR.	3
CUMMULATIVE # IN PRESS		BOOKS 0	PAPERS	26	ABSTR.	1

II.C. Resource Summary Table

The Resource Summary Table includes the totals from the previous sections of *Core Research and Development* and *Training*. The totals for *Collaborative Research and Service* we derived as the sum of the following components:

- **Usage Factor**. These are computed as the differences between the totals for the BIONET minus the Core R+D and Training figures. The total for *CPU Min. Used* is the sum of all CPU time consumed by BIONET users and staff. The total of staff hours is self explanatory.

- **BRTP Funds Allocated**. This is computed as the difference between the total budget for BIONET minus the categories of Core R+D and Training, and minus the capital equipment expenditures for the year ($8000).

The category of *Collaborative Research and Service* includes an entry of $153,100 in the column *Other Funds*. This is the total money collected over the period 12/85 - 11/86 for subscription fees. Each PI is asked to pay an access fee to help defray the telecommunication costs for access to BIONET; this fee is currently $400/year. By agreement with BRTP, these access fees are not grant related income.

The balance of these fees carried forward from the previous year (as of 12/1/85) was $25,000. After twelve additional months of collecting subscription fees and disbursing them for telecommunication expenses, the balance is now $120,100. BIONET has increased the number of telecommunication ports by four to provide improved service. This action has increased the telecomunication charges by a factor of two which should result in a total expenditure for telecommunications of approximately $200,000 during the next twelve months. At this rate we have judiciously reserved the current balance to serve our needs over the next calender year and plan on substantially depleting this current balance by November of 1987 in accordance with the timing of the receipt of the subscription fees from individual users.

The category of *Administration/Miscellaneous* includes only the Usage Factor of BIONET's share of the DEC-2060 system overhead accounts. No facility staff computer time, work hours, or BRTP funds are allocated; we consider such time and funds to be an integral part of the support of the other components of the Resource. We do include as Funds Allocated the $8000 to purchase items of capital equipment.

The category of down time includes the sum of scheduled and unscheduled maintenance on the DEC-2060 computer. In the period 12/85 - 11/86, there was a total of 99 hours (5935 cpu minutes) of downtime:

- 3266 cpu minutes of scheduled downtime for preventive maintenance and several system-related tasks including major time for installation of Version 6.1 of the TOPS-20 operating system.

- 2669 minutes of downtime were due to unscheduled maintenance.

The downtime reported in the Summary Table is 50% of the total, reflecting BIONET's access to 50% of

the machine. Note that the (total) unscheduled maintenance of 2669 minutes is only 0.5% of the total cpu time available. Considering both categories of downtime, the machine has been available for use by BIONET scientists, 99.5% of the time, 24 hours a day, seven days a week. No funds have been allocated to this category.

PART II, SECTION C RESOURCE SUMMARY TABLE

AWARD NUMBER 5U41RR01685-03 REPORT PERIOD March 1, 1986 to February 28, 1987

RESOURCE COMPONENT	(1) Number Subproject	(2) Number Publications	(3) Number Investigators	(4) USAGE FACTOR Resource a/ Technology	CPU Min Used	Resource Staff Hrs	(5) BRTP Funds Allocated $	(6) Resource Fees $ Collected	(7) Other Funds $
CORE RESEARCH & DEVELOPMENT	3	-0-	Staff as Above	DEC-2060	1810	1176	83,385		
COLLABORATIVE RESEARCH & SERVICE	489	114	1709	DEC-2060	7990	9283	538,235		153,100
TRAINING	N/A	N/A	Staff as Above	DEC-2060	664	664	41,012		
ADMINISTRATION/ MISCELLANEOUS	SEE PREFACE TO THIS		SECTION	DEC-2060	8466		8,000		
DOWN TIME					2968				
GRAND TOTALS	492	114	1709		24,845	11,123	670,632		

a/ Identify Resource Technologies Used. b/ Give Hours Resource Technologies Used. See Instructions, page iv.

Part III. Narrative Description

III.A. Summary of Research Progress

We have had a very successful year of operation. Previous problems in funding were resolved, and our operation has stabilized to the great benefit of the scientific community. Although our funds were cut 8% at the beginning of the year, we were able to anticipate those cuts and our budget was supplemented in October 1986 by funds transferred to the NIH from the National Science Foundation to help support telecommunications. This support, plus the continuing support from our Principal Investigators in the form of subscription fees for access to BIONET, has allowed us to focus on the scientific goals of the Resource rather than on fiscal dislocations. As a result, we have been able to provide high quality, uninterrupted service to the BIONET community and devote additional personnel resources to Collaborative and Core Research projects. A possible cloud on the horizon is a lack of resolution of the indirect cost rate. This could seriously affect our future performance.

The following sections describe in detail our accomplishments in the several components of the BIONET Resource. Here, in brief, are some of the most notable:

- The community in terms of active PI's has grown substantially in the past year. To serve better this community, we have increased the number of communication ports into BIONET. The statistics on the community were described in the introduction to Chapter II. The new communication facilities are described under *Resource Facilities*, below;

- The community published or has in press more than 114 papers during the past twelve months, an indicator of the crucial role BIONET plays in computational support of research in molecular biology;

- The response of the central DEC-2060 computer during prime time has become quite slow, due to the high level of use by the community. We are addressing this problem through installation of additional BIONET Satellites and through mechanisms for distributing time consuming computations to other machines (see *Core Research* below).

- We have taken several steps to improve the electronic communications available through BIONET, including substantial work on revamping the bulletin board system (see *Collaborative Research* below), and installation of mail forwarding among facilities via Telenet and ARPANET (see *Core Research* and *Resource Facilities* below);

- We have begun the collection and dissemination of several programs designed to help solve the very important problem of alignment of multiple biological sequences. This important Core Research activity will place BIONET in a leadership role in making such programs routinely available (see *Core Research* below).

- We have identified several vendors for special text searching hardware and identified two as providing the most promising machines. Such machines have the potential for revolutionizing methods for text search and calculation of sequence homologies. We are currently making arrangements for access to these machines (see *Core Research* below).

III.A.1. Service

The Service component of BIONET includes primarily Class I investigators who use the BIONET Core and Contributed program Libraries (see section III.A.5.c) to support their research. They have access to all functions of the System Library (section III.A.5.d) as well, but use primarily the systems for electronic communication (electronic mail and bulletin boards).

The Service component also includes Class III investigators who are given access based on their responsibilities for computing in molecular biology using department, school or campus-wide computer facilities. As part of their agreement for membership in BIONET, they provide information about the resource and access to it for their local community. There are currently four investigators in this category:

- Robert Gross - Dr. Gross has responsibilities in the newly-created Molecular Genetics Center at Dartmouth. He facilitates local use of BIONET, and has helped arrange a training at Dartmouth (see *BIONET Training Program* below);

- Kenneth Manly - Dr. Manly is a Cancer Research Specialist at the Roswell Park Memorial Institute at SUNY-Buffalo. He provides local support for BIONET users at this institution;

- Pavel Vitek - Dr. Vitek is Head of Computer Facilities at the Imperial Cancer Research Fund in London. He provides local support and serves as a source of information on BIONET for other UK investigators;

- Charles Lawrence - Dr. Lawrence is Principal Investigator for the Molecular Biology Information Resource, a regional resource funded by BRTP/DRR/NIH. He provides information about BIONET to the regional community supported by his Resource.

The final category of Service users is Class IV, a Class provided for those investigators who wish access only to the communication facilities of BIONET. There are currently four investigators in Class IV: Marlene Belfort, New York State Dept. of Health at Albany; Dieter Soll, Yale University; Jean Walat, BIOSIS; and Edward Hoover, Colorado State University.

III.A.1.a. Scientific Consulting: Class I, III and IV Support

The Service component of the BIONET Resource is supported by a group of BIONET Scientific Consultants. The Consultants interact with the community in a variety of ways, including direct support via telephone calls, electronic mail and terminal links with individual investigators. Support is also provided through their participation at major meetings, trade shows, and trainings. The consultants also provide on-line and printed documentation for User Manuals, program descriptions and system procedures.

We currently have two full-time and one half-time Consultants. They provide direct support to the community 50% of their time. The other 50% is devoted to participation in Core and Collaborative Research projects as described in subsequent sections.

III.A.1.b. Service

The Service component of the BIONET Resource includes primarily Class I investigators and takes the form of answering questions by phone, by electronic mail, and by terminal links. A survey of the monthly phone, mail and terminal links for the past year shows the different uses of the BIONET Resource.

The monthly inquiry rates for the five categories of Programs and Databases, TOPS20 System, PCs and PC Software, Telecommunications, and BIONET Administration are listed below.

Table III-1: Summary of Monthly Rates of Inquiries

Category	Number of Inquiries	Percent of Total Inquiries
Programs and Databases	153	36
TOPS20 System	117	27
Telecommunications	72	17
Administration	45	10
PC and PC Software	41	10
	----	----
TOTALS	428	100

As can be seen from Table III-1, the largest number of inquiries concern the use of BIONET's programs. This first category, Programs and Databases, has been subdivided into five different scientific and program categories in Table III-2.

As shown in Table III-1, the largest number of questions received by the BIONET Staff concern the use of the resource's programs. The analysis in Table III-2 shows that the majority of these inquiries relate to the use of the databases and the database access programs. The sequence entry/manipulation and sequence analysis programs categories were a fairly distant second at 18% and 15% respectively, and the rest of the program categories accounted for less than 10% each of the total rate of inquiries. Multiple-Sequence alignment inquiries mostly concerned the use of William Bains' contributed XMULTAN program. TOPS20 system program inquiries mainly concerned the use of FIND and XSEARCH, and experiment planning and analysis inquiries covered the use of the SIZER, MAP and CLONER programs. The category OTHER includes inquiries on the use of other contributed programs, such as Michael Zuker's BIOFLD, and other miscellaneous questions.

Returning to Table III-2, 27% of user questions concerned the TOPS20 operating system, specifically the manipulation of files and directories, the use of the text editors, the control of output to the terminal, and the access of programs.

Table III-2: Summary of Monthly Rates of Questions for Programs and Databases.

Category	Number of Inquiries	Percent of Total Inquiries
Database Searches and Databases	65	42
Sequence and Gel Data Entry and Manipulation	27	18
DNA and Protein Sequence Analysis	23	15
Multi-Sequence Alignment	11	7
TOPS20 System Programs	10	7
Experiment Planning and Analysis	5	3
Other	12	8
	----	----
TOTALS	153	100

The third largest category in Table III-1 is Telecommunications. This category includes both inquiries concerning the procedures involved in connecting to the BIONET computer and the quality of the communications between remote users and BIONET.

The last two categories in Table III-1, BIONET Administration and PCs and PC software, each accounted for 10% of the inquiries. The BIONET administration category consisted mostly of application requests, training session information, and manual requests. The majority of the BIONET administration

calls were routed directly to the BIONET Administrator and are not included in Table III-1. The questions most frequently asked about PC software pertained to file transfer and terminal emulators. Questions of this nature mainly concerned software which runs on PCs manufactured by IBM and Apple.

The majority of the users' questions were answered immediately by our Scientific Consultants. Even the more difficult questions usually received a response within a day. The availability of the Scientific Consultant staff resulted in substantial savings of investigator research time.

III.A.1.c. Scientific Case Studies Using BIONET

After examining the large number of publications received, we chose the following two examples of research which utilized the BIONET resource.

"Evolution and High-Order Structure of Architectural Proteins in Silkmoth Chorion" EMBO Journal 5:1981-1989, 1986, J.C. Regier.

Jerome C. Regier joined BIONET in February 1985 and has used over 1400 CPU minutes and 158 connect hours during the past year on BIONET.

Dr. Regier has been studying the structure of silkmoth chorion with the aim of understanding eukaryotic morphogenesis at the molecular level. Silkmoth chorion morphogenesis involves the temporally regulated production of more than 100 follicle cell-specific proteins in widely varying amounts. The chorion proteins assemble to form two predominant types of structures: lamellae that are highly ordered helicoidal arrays of fibrils and filler that is a sponge-like network. Filler accounts for roughly 5% of the chorion's mass and consists of only two proteins E1 and E2. It forms hollow breathing channels through lamellar chorion and molds a small number of outer surface lamellae into crown-shaped structures called aeropyle crowns. Prior to this report, the E1 cDNA and genomic clones had been sequenced. Their predicted secondary structures and hydropathicity profiles revealed a periodicity postulated to have functional significance. No homology was detected between E1 and lamellar sequences.

Dr. Regier's group sequenced genomic and cDNA clones that encode the E2 silkmoth chorion protein. E2 was found to have two distinct domains with the amino terminal domain consisting of four alternating stretches of hydrophobic and hydrophilic residues, the first three of which are homologous in sequence to about half of the E1 protein. The carboxyl terminal domain of E2 is much longer. It is hydrophilic and consists entirely of multiple tandem copies of a single, variant hexapeptide repeat sequence that is absent from E1.

Dr. Regier searched for homology between the nonrepetitive region of E2, which spans residues 1-127, and the sequences found in the three large databases that are available on BIONET: the National Institutes of

Health DNA sequence library (GenBank), the European Moleculary Biology Laboratory DNA sequence library, and the National Biomedial Research Foundation's protein sequence database. This search was performed using the IFIND program. E1 gave the highest similarity score of any sequence in all three databases and was the only sequence whose score did not form part of the continuum of other scores.

He also searched for similar lysine- and asparagine-like hexapeptide repeats in other sequences, using BIONET's QUEST program. For the protein searches, he looked for at least three tandem repeats of "Lys Lys Asp ..." or four of "Asn .....", where a "." represents any amino acid. None were found. For the nucleotide searches, he looked for coding regions that contained at least three tandem repeats of "AA.AA.GA. ... ..." or four of "AA[T or C] ... ..." For the first octadecanucleotide, no sequences were found. For the second, four groups of closely related sequences were found that could be translated in the appropriate open reading frame. One (plasmodium surface antigen) encoded multiple tetrapeptide repeats of Asn-Ala-Asn-Pro. The second (trypanosome ribosomal protein) contained long stretches of polyasparagine. The last two (drosphila GAG-like protein and yeast positive regulatory protein) contained no evidence of being within a repeat region other than for the presence of the four asparagine codons. He concludes that the lysine and asparagine repeats of E2 have not been found in other sequences.

Based on the above sequence information, he suggests that the most likely evolution for the E genes is that the ancestral E gene encoded an amino terminal sequence similar to that in the modern-day E1 and E2. After the ancestral gene duplicated *in toto*, one of the copies (the E2 gene ancestor) added on the hexapeptide repeat domain. This could have occurred either by reduplicating a sequence already present within the gene or by transposition of an unrelated sequence. He theorizes that the absence of an intron at the border of the repeat domain argues against the exon shuffling hypothesis. They are currently sequencing E genes from another silkmoth species in hopes of better understanding the origin of the repeat domain.

"Intron Mutations Affect Splicing of *Saccharomyces Cerevisiae* SUP53 Precursor tRNA", Molecular and Cellular Biology, 6: 2674-2683, 1986, M.C. Strobel and J. Abelson.

John Abelson has been a BIONET user since December 1984 and has used over 600 CPU minutes and 206 connect hours of BIONET computer time in the past year. The Abelson lab also actively uses BIONET's electronic mail and bulletin board facilities to further their research efforts.

Drs. Strobel and Abelson investigated the role of intron structure and sequence on precursor tRNA splicing and subsequent mature tRNA production. In their experiments, they used mutants of the *saccharomyces cervisiae* amber suppressor tRNA gene SUP53 that encodes a pre-tRNA containing a 32 base intervening sequence. They constructed two types of mutants: a first type with an internal deletion

of the natural SUP53 intron and a second type with a novel intron. These mutant genes were transcribed *in vitro*, and the end-processed transcripts were analyzed for their ability to serve as substrates for the partially purified *S. cerevisiae* tRNA endonuclease and ligase. After integration of these mutant genes into the yeast genome, the *in vivo* suppressor tRNA function of these mutant alleles was correlated with the *in vitro* phenotype.

The *S. cerevisiae* pre-tRNA introns have a common structure where each intron folds to form an extension of the anticodon stem. The splice junctions are generally in single-stranded loops and are a constant distance apart (6 base pairs). The extended anticodon stem is stabilized by base-pairing between the anticodon and the intron, allowing the stem nucleotides to be highly base-stacked while the loop nucleotides are not.

The secondary structures of SUP53 and the mutant RNAs were determined by using the RNA sequence folding program of Michael Zuker which is available on BIONET. These structures were determined both as part of the total pre-tRNA and as subfragments spanning the anticodon stem-loop region (including intervening sequences).

Drs. Strobel and Abelson's experiments showed that certain intron-mutated precursor tRNAs were refractory to the tRNA-splicing endonuclease both *in vitro* and *in vivo*. Furthermore, secondary mutations at the 3' intron-exon junction could partially rescue the nonspliced phenotype *in vitro* and *in vivo*.

Single-nucleotide changes were shown to have a drastic effect on the predicted secondary structure of the intron. The results were consistent with the idea that correct, efficient splicing is dependent on the intron's ability to form a stable extension of the anticodon stem and retain the 3' splice junction in a single-stranded region. The BIOFLD program on BIONET was used to predict the secondary structure of each precursor. Precursors whose secondary structures formed a stable extended stem were readily spliced. In contrast, two unspliced precursors exhibited severe distortion of the extended stem. Additionally, while tRNA precursors exhibited no sequence conservation at exon-intron junctions or within introns, certain primary sequences at these junctions were thought to be inappropriate.

III.A.1.d. Comments from the Community

A measure of the success of BIONET is the response from the community about the Resource. As part of our reapplication procedures for renewal of access to BIONET, we include a section for comments and suggestions. These comments and suggestions are a valuable source of information to us in order to improve BIONET. In this section we summarize the comments and suggestions received during the current cycle of renewals (timed to coincide with preparation of our Annual Report). See Appendix I for a copy of the reapplication form.

It is clear from the comments received that BIONET is a valuable part of all research efforts entailing nucleic acid or protein sequence analysis. The most valued aspect of the computer resource is the ability to search the major databases. Fifty five percent of those responding cited this reason. The most commonly used search is for sequence homologies. Other capabilities cited included the usefulness of the programs for the assembly and analysis of sequence data. Although the majority of the work on BIONET involves nucleic acid sequence analysis, it is interesting that about 25% of the citations specifically mentioned the use of BIONET for protein sequence analysis. In addition, it was gratifying to note that 15% of the repondents cited the usefulness of the bulletin boards in their research efforts. If the citations concerning the electronic mail system are included, approximately one quarter of the respondents mentioned the use of the BIONET communications facilities. Although this usage is still below that of the sequence analysis programs on BIONET, one should keep in mind that bulletin boards and electronic mail are still rather novel concepts in the biological sciences. When viewed in this light, the figures show that the acceptance of electronic communications by the research community is off to a good start.

The comments also detail the need for further improvement of the BIONET resource in several areas. We have already begun to address these remaining points of user disatisfaction even before receiving the reapplication comments.

The main problem cited in the comments section remains the slow response time and heavy usage of the time-shared DEC 2060 system. We are taking several steps to address this problem. In addition to our efforts to reduce the user load through the BIONET Satellite Program (section III.A.3.b below), we took immediate action to encourage use of the batch job facility on the 2060. Batch jobs allow users to run CPU-intensive tasks during off-peak hours. Our efforts included (1) a thorough rewriting and simplification of the documentation describing the batch technique; (2) dissemination of information about the technique via numerous bulletin board notices and login banner messages; and (3) individual assistance offered to users by our consulting staff. In most instances the consultants wrote the first batch job control file for investigators who then proceeded to use it as a model for further efforts on their own.

The second most often cited complaint concerned the uneven quality of the program documentation. With the addition of new staff at BIONET we have already been able to address part of this problem through the addition of new on-line help menus (see section III.A.1.f below and Appendix II). We have also begun work on a new manual of examples. This should enable users to find rapidly the appropriate sequence of commands required to accomplish their research tasks. These examples will first be presented on-line to obtain user feedback and will then be made available in printed format after they meet with user approval. The first of these examples will go on-line during the month of January 1987.

The recent switch to the TELENET telecommunications network caused some degree of confusion and

dissatisfaction, but we have taken action to address individual phone connection problems as they have arisen. As a result the number of problems of this nature is diminishing (see section III.A.5.a)

Several investigators expressed concern about the extended period of time between the publication date of new sequence information and its availability in the GenBank and EMBL databases. We are developing new procedures to allow BIONET users to contribute their sequences for on-line use as soon as the data is ready for submission to the major databanks. This project is described in section III.A.2.d below and should commence during January 1987.

The remainder of the comments involved individual suggestions for additional features on the resource and we are considering each suggestion carefully. Several of these requests have already been implemented. These recent changes include the addition of the Brookhaven protein databank and better editors on the system. Requests for more graphical program output, and dot matrix homology search programs are in the process of being addressed.

Finally, we note that only two investigators out of the 171 reapplicants complained about the cost of the subscription fee indicating that this is no longer a major area of concern.

III.A.1.e. PC/BIONET Communications - Distribution of the Resource

It has been clear from the beginning of the operation of BIONET that the majority of the user community had access to personal computers, and that they were looking for ways to use the PC's effectively in conjunction with BIONET. For example, sequences may be entered using a PC and stored on a floppy disk without the need to connect to BIONET. The file can be transmitted to BIONET later over the telephone lines. This reduces the load on the system. We have strongly supported this method of access, to the extent of maintaining a lending (and on-line) library of software and documentation for file transfer and terminal emulation programs. The growing availability of PC-based software for sequence analysis is another means by which the burden on the DEC-2060 will be reduced significantly.

To further facilitate the use of PCs, the IBM-PC public domain version of BIONET's on-line EMACS editor has recently been promoted on the system. "MicroEMACS" is distributed free to users and has the virtues of producing ASCII text files compatible with the BIONET software and of utilizing essentially the same command set as the mainframe editor. This should facilitate sequence entry on user PC's. The consultants have also provided assistance in converting files into a format compatible with the on-line software. They are actively encouraging long-time BIONET users to switch from the older SOS line editor to EMACS. The availability of an on-line interactive tutorial for learning EMACS, as well as recent improvements in on-line documentation for both EMACS and PC-to-mainframe file transfers, should promote this change.

BIONET has also sought to standardize file transfer protocols on the system by vigorously promoting the use of the public-domain Kermit software. Several tests conducted by the BIONET staff during the transition from UNINET to TELENET proved once again that Kermit was the most trouble-free of the file transfer protocols. An installation program has been added to the IBM-PC version of Kermit in the BIONET lending library. This assists scientists in their initial use of the software.

III.A.1.f. System Help Utilities

To facilitate use of BIONET, information about the system has been organized in a series of menus which users can view by typing "HELP ME" at the system level. A copy of the main menu accessed by this command is provided in Appendix II. The topics listed on this menu include (1) a guide to the IntelliGenetics programs and contributed software on the system; (2) instructions for using the electronic mail and bulletin board systems; (3) how to locate other users on the system; (4) operating system help; (5) on-line database information; (6) batch job instructions; (7) how to obtain consulting help; (8) how to use the on-line text editors; (9) how to transfer files between microcomputers and mainframes; (10) and assistance for TELENET. An upcoming feature on the main menu will be an index which will refer to new on-line program usage examples.

Several of the menu topics are broken down further into subtopics in lower-level menus (see Appendix II). This organization allows the user to remember only a single command, "HELP ME", to find his/her way around the system. This command is reinforced continually by a login banner reminder. After the program examples are implemented, we anticipate that any remaining difficulties in learning the use of the BIONET resource will be completely resolved.

In the few months that the HELP ME menu has been available, it has rapidly become the most frequently used help file on the system, being accessed at an average rate of just over 300 times a month. The consulting staff has received numerous favorable comments about the menu system during telephone conversations with users and also in electronic mail messages. With the coming addition of on-line program examples, HELP ME should become even more useful to the BIONET community.

III.A.2. Collaborative Research

BIONET's collaborative community is made up of several components, encompassing efforts by outside scientists working in conjunction with BIONET staff. In subsequent paragraphs we discuss each component in more detail:

- **DEC-2060 Software Contributors.** This component includes those persons who have contributed software for use by the BIONET community on the central DEC-2060 computer;

- **PC-Based Software.** This component includes our efforts to gather and disseminate PC-based software of special utility to the community.

- **Data Contributors**. This component includes those persons who, together with BIONET staff, contribute data useful to the community;

- **Liaison with Other Resources**. Several accounts have been established to promote sharing of information among molecular biology computing resources;

- **Bulletin Boards**. This component involves scientists who have agreed to maintain bulletin boards of special interest to the BIONET community.

III.A.2.a. DEC-2060 Software Contributors

The primary efforts of the collaborative research this year have gone into helping make programs developed by BIONET users available to others. This has included (1) conversion of a number of programs from other languages and operating systems to the BIONET environment; (2) altering the programs to accept BIONET file formats; (3) setting up bulletin boards announcing programs available for downloading to microcomputers; and (4) preparing lending libraries of the programs which can be physically sent out to BIONET users.

The following sections describe each of the programs that we have made available on BIONET, either as a program on the BIONET computer or as a program that can be executed on a local microcomputer. For programs that are used on BIONET we can provide a direct measure of their use. For those programs used on microcomputers we can only provide the number of times that the program file has been referenced (downloaded) and the number of independent requests for a copy of the program via the lending library.

DFASTP and DFASTN - BIONET and IntelliGenetics have made substantial efforts to bring up the DFASTP and DFASTN programs from Dr. Bill Pearson and to make them available to the entire community. IntelliGenetics ported the code for Pearson's programs from his VAX version to TOPS-20 versions (XFASTP and XFASTN) at no cost to BIONET. These programs are used heavily to search the PIR protein or the NIH GenBank databanks for homology with newly determined sequences. XFASTN and XFASTP are substantially faster than the IFIND program (which is based on the original Wilbur and Lipman NUCALN algorithm) and help reduce the total amount of computer time that BIONET uses in such searches.

BIONET users have performed 373 searches of the current PIR protein database using XFASTP. This amounts to 97 searches per month or over 3 searches per day. Unfortunately this is far less than the use of IFIND by BIONET users (346 searches per month or over 10 searches per day) searching the same database. The current BIONET documentation describes how to use the IFIND program but the use of the contributed XFASTP is only documented on line and in bulletin board messages. We can stimulate use of XFASTP by further messages on BBOARDS. Since IntelliGenetics has licensed the XFASTP and XFASTN programs from Dr. Pearson for incorporation into their IFIND program, the difference between

CPU usage of IFIND and XFASTP will be eliminated upon the release of the new versions of the core software.

XMULTAN - XMULTAN is a program developed by Dr. Bill Bains for aligning multiple homologous DNA sequences. While it can only align sequences which are at least 60% homologous, it is an extremely rapid program and expands the capabilities of the resource. Multiple sequence alignment is useful for BIONET users studying evolution and for those trying to obtain a consensus from many sequences of similar function. Appendix III shows seven related satellite DNA sequences isolated from several sibling species of *Drosophila* which were aligned with XMULTAN in less than 1 minute of CPU time on BIONET.

XMULTAN was modified by BIONET to be compatible with our file system and a menu driven front end was added to make the program easier to use. The original version was a non-interactive batch oriented program. A number of BIONET users have successfully used XMULTAN and provided us with valuable feedback. Although available for only two months, it is currently being used at a rate of 65 times a month.

RNAFOLD and BIOFLD - Last year Dr. Michael Zuker, from the NRC laboratory in Ottawa Ontario made BIOFLD available as a program on the BIONET computer. This program predicts RNA secondary structures and has been used 509 times for an average rate of 24 times per month.

ALIGN - Dr. Dan Davison, while a graduate student at Stony Brook at SUNY wrote and contributed two versions of his ALIGN program. The first version runs directly on the BIONET computer. It can be used to align two very long DNA sequences including those which are not very homologous to each other and contain large gaps. The alignments are significantly better than similar heuristic alignments obtained from the SEQ SEARCH procedure in the BIONET core programs. The alignments are not as good as obtained from the SEQ ALIGN procedure, but Dr. Davison's ALIGN program is significantly faster. The ALIGN program has been used 155 times (27 times per month average).

IDEAS - While at the NIH Dr. Kanehisa contributed this package of software to BIONET and made the program compatible with ours. The usage of this package has decreased this year to about once per month for two reasons. First, much of the functionality of these programs is already contained within the IntelliGenetics core library, and secondly, Dr. Kanehisa is now at Kyoto University and has interacted less with the resource due to communications costs.

XPROF - Dr. George Rose at Pennsylvania State University has contributed the DEC-VAX Fortran version of his method for calculating hydropathicity profiles for proteins based on empirical observations on the extent to which amino acid residues are found to be exposed or buried. We have modified this program to run on the DEC-2060.

A number of new collaborations are currently underway to bring even more functionality to the BIONET system. This includes obtaining a number of programs requested by BIONET users as well as a number of new relevant databases.

New Multiple Sequence Alignment Programs - XMULTAN is primarily useful for alignment of nucleic acid sequences since homologous proteins are generally less than 60% identical. A number of programs have been described in the literature for performing multiple alignments of protein sequences. Dr. Joel Sussman of the Weizmann Institute has made available his VAX Fortran program for aligning three protein sequences simultaneously. This program is a straightforward extension of the Needleman-Wunsch method for aligning two protein sequences and is currently being converted for use on BIONET. Unfortunately, it will be limited to aligning only three sequences at a time, and like the original Needleman-Wunsch procedure, it is a CPU and memory intensive application.

IntelliGenetics is also developing a new program for multiple sequence alignment of proteins and DNA sequences in collaboration with Dr. Hugo Martinez of the University of California, San Francisco. This program (GENALIGN) will be available in January of 1987. GENALIGN is a very rapid program for multiple alignment of either protein or DNA sequences and should adequately fill the need for both multiple protein sequence alignment and for consensus DNA alignments at low levels of homology.

Phylogenies And Evolutionary Trees - A suite of programs for constructing phylogenies based on sequence relationships has been written by Dr. Joe Felsenstein, Department of Genetics, University of Washington. Dr. Felsenstein has made these programs, originally written for a VAX computer, available to us. BIONET is currently modifying them for use on our resource. This program, coupled with the multiple sequence alignment programs mentioned above will provide the essential tools to carry out evolutionary biology studies with BIONET.

Finally, we expect that connection to the ARPANET will make a large number of libraries of public domain software readily available to BIONET users. One of the IntelliGenetics staff already maintains a collection of Macintosh public domain software obtained largely from the SUMEX computer resource at Stanford University. With the pending connection of the BIONET computer to the ARPANET the following software collections will be also available: <INFO-IBMPC> at ISI; <INFO-KERMIT> at COLUMBIA; <INFO-MAC> at SUMEX; <PC-BLUE>, <CPM>, <UNIX> and <MS-DOS> all at SIMTEL-20, <IBM> at DEC Marlboro etc. We will support access to these compilations of programs by publishing bulletins and procedures for accessing these computers via file transfer protocols. We will also maintain lists of programs available at each of these sites. With the advent of these resources directly available on ARPANET we will probably stop local maintenance of the <KERMIT> and <MACINTOSH> directories.

III.A.2.b. PC-Based Software

We began the lending library concept last year by making Kermit (a public domain terminal emulator and file transfer protocol from Columbia University) available to BIONET.We send them a diskette and documentation which they copied and returned to us. We currently make Kermit available for Apple II, Macintosh, IBM-PC and TRS-80 model computers. This year we have extended our lending library to include many BIONET-user developed programs and a number of utility programs that are useful for file transfer on IBM-PC and Macintosh computers. A complete list of the programs which are available is in Appendix IV. The <PC-SOFTWARE.MAC> and <PC-SOFTWARE.IBM> directories in particular have been referenced an average of 17 times a month each. In addition we have sent out diskettes 32 times that contain these utility programs. Their primary functions are to allow users to share software by helping them to upload and download files from BIONET and other mainframe computers.

We have also set up bulletin boards concerning PC-SOFTWARE in addition to PC-COMMUNICATIONS and these have been among the most popular bulletin boards with 107 and 47 message postings respectively last year. A number of bulletins give detailed procedures for performing electronic transfer of both text and binary files.

RNAFOLD AND BIOFLD - This year Dr. Zuker developed an IBM-PC version of his mainframe program. When this version is run on an IBM AT with 640 K of memory it is capable of calculating the minimum free energy secondary structure of RNAs' up to 360 bases long. His program has been downloaded from BIONET 51 times and in addition we have sent out copies of this program from the lending library.

MOLECULE - One of the weak aspects of Zuker's BIOFLD and RNAFOLD program is the display of the predicted secondary structure of the RNA. Both versions print a one dimensional representation of these inherently two dimensional structures. Dr. John Thompson (a post-doc in Professor John Woolford's group at Carnegie Mellon University) has written a program which displays the optimal secondary structure in two dimensions on a graphics screen. The MOLECULE program uses the connection output file (CT file) that is produced by both Zuker's RNAFOLD and BIOFLD programs. It may also utilize files produced by the NUCSHO program developed at NIH by Richard Feldemann. All these programs for displaying RNA secondary structure are available on BIONET. Appendix V shows a graphics display of a minimal energy secondary structure for a viroid RNA molecule as presented by the MOLECULE program. This program has been downloaded 51 times (average of 10 times per month) and in addition we have sent out copies of this program from the lending library.

ALIGN - Dr. Dan Davison has made an IBM-PC version of the program available for people to download and for our lending library. The IBM-PC version of ALIGN has been downloaded 20 times.

DM - Drs. Bruce Conrad and David Mount have developed a significant library of DNA analysis programs for use on IBM-PCs. They have made these programs available through a distributor who charges handling fees only. They also volunteered to place them in our public domain directories. BIONET users have downloaded this set of programs 31 times (about 5 times a month on the average). This is quite impressive usage considering the size of these programs (362K). These programs carry out editing, display and plotting of both DNA and protein sequences. They also have a complete restriction enzyme analysis package. They can plot dot matrix comparisons between sequences and have a very sophisticated oligonucleotide search capability. This is one of the most comprehensive software packages for sequence analysis in the public domain and we are very grateful to Dr. David Mount for making it available to the BIONET facility.

BIONET has always provided rapid updates to all the major collections of sequence data. These include the GenBank and EMBL nucleotide sequence collections and the NBRF/PIR Protein Data Bank. There are many other types of data of use to the community, however. Recently, we have been encouraged by Dr. Rich Roberts of our National Advisory Committee to view as legitimate Collaborative Research the collection and dissemination of various data sets related to molecular biology. The following summarizes our current and projected activities in this area:

- **Restriction Enzyme Database**. We continue to provide the community with the latest additions to the Restriction Enzyme Database, through the cooperation of BIONET and Dr. Roberts at Cold Spring Harbor (CSH). As described in our last Annual Report, modifications are uploaded automatically to BIONET shortly after they are incorporated into the on-line database at CSH. In addition, we provide the community with subsets of the list of enzymes that are commercially available.

- **VectorBank^{tm}**. IntelliGenetics has made substantial additions to its databank of map and sequence data for plasmids. This will shortly be released as part of updates to the Core Library, and thus will be available to BIONET. We have also urged the inclusion of the EMBL plasmid sequences in the EMBL database.We will add these to VectorBank as soon as they are available.

- **GNOMIC**. The publication *GNOMIC, A Dictionary of Genetic Codes*, by E.N. Trifonov and V. Brendel is an indexed compilation of important short nucleotide sequences. There are many entries in this compendium that may be used in database searches. We plan to enter these sequences, with annotations, into a format compatible with the *QUEST* program for pattern searching in biological databases. These will be made available as new "key" files which are used directly in *QUEST*.

- **New Sequences**. We will shortly establish a directory to which scientists can contribute new sequence data for examination by others prior to their formal acceptance and inclusion in the major databases. The <NEW-SEQUENCES> directory should be ready for user contributions around the end of December 1986 and a SEQUENCES bulletin board may be implemented to announce contributions. Sequences that have been accepted for publication will be differentiated from other contributions to the directory, and a simple index scheme will be provided for files in the directory. The current plan is to leave the responsibility for data quality control with the sequence contributor, and users will be notified to this effect. This

effort may be expanded depending upon user response and the availability of BIONET staff time.

III.A.2.c. Liaison with Other Resources

Several accounts have been established on BIONET to promote interaction with other related Resources. The following is a summary of current sites with which we can exchange information, together with a brief list of sites that will be accessible when the full implementation of ARPANET gateways is established:

- **Molecular Biology Computer Research Resource.** The MBCRR, at the Dana-Farber Cancer Institute at Harvard, shares information through mail delivery via the GENE account and via the bulletin board system. Most recently, these facilities were used heavily in support of the New Hampshire meeting on *Macromolecules, Genes and Computers* sponsored by the MBCRR. The MBCRR also supplies BIONET with a formatted version of the Protein Data Bank, organized along functional lines, and we have made this available to the BIONET community;

- **Molecular Biology Information Resource.** The MBIR at Baylor College of Medicine communicates with BIONET through Dr. Lawrence's account for electronic communication.

- **Protein Identification Resource.** Information exchange with the PIR at the National Biomedical Research Foundation is through the account of Winona Barker, the Principal Investigator;

- **GenBank.** Information exchange with GenBank is via accounts set up for communication with both Bolt Beranek and Newman and Los Alamos National Laboratory. The former is via Dr. Howard Bilofsky's account. The latter, LANL, is via accounts set up for Dr. Walter Goad, with sub-accounts for key personnel on his staff. This interaction is more intensive due to the importance of information exchange on new sequence data and through their participation in the GenBank bulletin board and related activities (see the next section on bulletin boards). A prototype of an automated sequence submission utility has been developed and tested at BIONET during the last year. This program will allow users to log onto BIONET and use its facilities to input a sequence for transmission to GENBANK. It provides the style of interaction familiar to BIONET users, and permits them to use any existing data file with which they have been working on BIONET. The goal of this facility is to make it easier and faster to submit sequences.

- **Other Resources.** Our impending connection to ARPANET will make communication with international resources substantially easier. All such resources have access to ARPANET via gateways to JANET in the United Kingdom, EARN in Europe, ACSNET in Australia, and BITNET equivalents in other countries. Thus, access will soon be simple to the European Molecular Biology Laboratory, the Molecular Biology Information Service in Australia, and the Imperial Cancer Research Fund in England. We already exchange information with the European Molecular Biology Laboratory through a temporary link routed through the SUMEX-AIM Resource at Stanford University.

III.A.2.d. Bulletin Boards

The following bulletin board topics are currently available on the system.

Bulletin Board Name	Description
ASK-BIONET	User queries and consultant responses
BION	Information for BION workstation users
BIONET-NEWS	General BIONET announcements
CONTRIBUTED-SOFTWARE	Information on programs contributed by users
EMPLOYMENT	Job openings
GENBANK	New bulletin board in preparation for GenBank information and queries
GENE-EXPRESSION	Scientific interest group
GENOMIC-ORGANIZATION	Scientific interest group
INFO-1100	Computer interest group
INFO-AMIGA	Computer interest group
INFO-ATARI16	Computer interest group
INFO-IBM-PC	Computer interest group
INFO-KCC	Computer interest group
INFO-KERMIT	Computer interest group
INFO-LAW	Assorted legal information
INFO-MAC	Computer interest group
INFO-SUN-SPOTS	Computer interest group
INFO-VAX	Computer interest group
METHODS-AND-REAGENTS	For reagent exchanges and announcements about lab methods
MOLECULAR-EVOLUTION	Scientific interest group
ONCOGENES	Scientific interest group
PC-COMMUNICATIONS	Information on communications software
PC-SOFTWARE	General PC software announcements
PLANT-MOLECULAR-BIOLOGY	Scientific interest group
POLITICS	Nonpartisan activities
PROTEIN-ANALYSIS	Scientific interest group
YEAST-GENETICS	Scientific interest group

The leaders of the individual boards are:

ASK-BIONET	Dr. David Kristofferson
BION	Jaya Carl
BIONET-NEWS	Dr. David Kristofferson
CONTRIBUTED-SOFTWARE	(to be named)
EMPLOYMENT	Mary Warner
GENBANK	Dr. Christian Burks
GENE-EXPRESSION	Dr. William Sofer
GENOMIC ORGANIZATION	Drs. Robert Jones & Stephen Harris
METHODS-AND-REAGENTS	Dr. Larry Kedes
MOLECULAR-EVOLUTION	Dr. Dan Davison
ONCOGENES	Dr. David Steffen
PC-COMMUNICATIONS	Mary Warner
PC-SOFTWARE	Dr. Doug Brutlag
PLANT-MOLECULAR-BIOLOGY	Dr. Robert Jones
POLITICS	Dr. Michele Cimbala
PROTEIN-ANALYSIS	Amos Bairoch
YEAST-GENETICS	Dr. John Thompson

Note: All the INFO- bulletin board material is received from information sources outside of BIONET.

Several changes were implemented in the bulletin board system during the final months of 1986. Four older bulletin boards (IMMUNOLOGY, LIBRARIES, VECTORS, and LAB-METHODS) were combined into a METHODS-AND-REAGENTS board. This increases the total number of subscribers who see each message and as a result should facilitate exchange of information and materials. Over 130 investigators on the system now subscribe to this new bulletin board.

A new ASK-BIONET bulletin board was started for the purpose of posting user questions and consultant answers that are of interest to all BIONET users.

All of the previous bulletin boards were scrutinized for use and a few inactive boards were deleted. Enthusiastic new leaders were found for others and some new topics were introduced. Although BIONET had a policy of waiving the $400 subscription fee for bulletin board leaders, this policy was not previously advertised. Its announcement on the system stimulated additional inquiries. It is still too early to assess the results of all the recent changes, but the situation is being monitored carefully. A brief description of some of the new activity follows.

The ONCOGENES board has started anew under the leadership of Dr. David Steffen who is maintaining an up-to-date compendium of oncogenes on it. Dr. John Thompson has recently begun directing the YEAST-GENETICS board.

A new GENBANK bulletin board is being implemented. The board will be directed by Dr. Christian Burks at GenBank in Los Alamos and will serve to facilitate communication between BIONET users and GenBank.

Dr. Robert Jones is the new leader for the PLANT-MOLECULAR-BIOLOGY board. He also shares the leadership of the new GENOMIC-ORGANIZATION board with Dr. Stephen Harris.

Amos Bairoch, the developer of PC/GENE, is leading a new bulletin board on PROTEIN-ANALYSIS which will deal mainly with protein sequence analysis methods and protein sequence data banks, but is open for input on any matter concerning proteins.

III.A.3. Core Research

III.A.3.a. Multiple Sequence Alignment

We have recently started work on a large scale project to review all the presently developed multiple-sequence alignment software with the goal of making available to the BIONET community as many of these programs as possible. After having surveyed the community that is involved in this kind of software research and development, we approached all the major developers and asked both for detailed information about their programs and also whether they would be willing to donate them for possible use by the BIONET community. So far we have received programs from the following researchers: Bill Bains from the University of Bath, England; Osamu Gotoh from the Saitama Cancer Center Research Institute in Japan, Mark Johnson from the University of Southern California, Hugo Martinez from the University of California at San Francisco; and Joel Sussman from the Weizmann Institute in Israel. In addition, we are waiting for programs from Michael Waterman from the University of Southern California and Wayne Anderson from the University of Alberta, Canada. As we receive the above software, we will conduct in-depth testing of their performances and when we feel that the software would be useful to the BIONET community, will release it after any necessary revisions. We currently have running on the 2060 the first three of these programs: Bill Bain's MULTAN, Hugo Martinez's MALIGN, and Joel Sussman's PROT3.

Bill Bains' MULTAN program, (see W. Bains, NAR 14: 159-177), has been updated and now includes, among other things, a more generalized input file format, user-specified output files, and reformatted output. We released this updated version, XMULTAN, to the BIONET community at the end of September 1986 and approximately a month and a half after its release, there have been 233 runs of the program with a usage rate of 172 per month. This usage rate falls in the range of the rates for the IG library programs of between 89 and 2020 runs per month and is higher than for three of the IG programs: SIZER, MAP and CLONER. The feedback so far has been that XMULTAN is meeting the needs of the BIONET Community for a program that performs multiple sequence alignments and generates a consensus sequence. XMULTAN takes many DNA sequences (it easily handles 30) of up to 1500 base pairs in length and generates the alignment and subsequent consensus sequence. The algorithm for XMULTAN is heuristic, and the actual running time for sequences with a fairly high degree of similarity (approximately 75%) is reasonably short. They range from 1-2 minutes CPU for a small number of highly similar short sequences to 10 minutes CPU for a larger number of less similar long sequences. The consensus generation for 10 sequences, 400 base pairs long, and 82% similarity took 7.6 CPU minutes. Bill Bains is presently rewriting MULTAN to include improvements to the algorithm and we hope to provide his new version of MULTAN to the BIONET Community as well.

Joel Sussman's PROT3 program to align three protein sequences is based on an extension of the Needleman-Wunsch algorithm (see M. Murata, J.S. Richardson, and J.L. Sussman, PNAS 82:3073-3077,1986). The program requires a large amount of storage space to run and because of the

machine constraints of our DEC 2060 only three sequences of no larger than 60 amino acids will run. We are presently looking into ways around this situation. PROT3 was written for the VMS operating system and does not present a storage problem on VAX systems. Additionally, the performance of this program and the other programs that we have received (especially Gotoh's and Johnson's) need to be assessed before we make any final decisions.

III.A.3.b. BIONET Satellite Program

This program has the goal of distributing the BIONET Resource among computers throughout the academic community. At the same time we want to establish better communication links among BIONET, its Satellites, and other computing resources in molecular biology. We currently have Satellites established at the Salk Institute, at the US Department of Agriculture, and at Fort Dietrick (US Army RIID).

We are following two approaches to communication with other facilities-ARPANET and a dial phone line-based network that we are simply calling the BIONET Network for the moment. Significant development occurred on the the BIONET Network this year. We have installed software on the BIONET central resource that can communicate with other computers through the same terminal connections that serve scientists using the resource directly. Since the majority of the computers at current or potential Satellite sites are Digital Equipment Corporation VAXes running the VMS operating system, we have implemented communication software for this operating system. To augment the basic mail capability of VMS, we have also arranged for distribution of an improved mail delivery program to BIONET Satellites. The software is scheduled to be delivered to the first BIONET Satellite for testing in December 1986.

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

In our last Annual Report we discussed our investigations of different machines designed for high speed searching of text for patterns of strings. We identified one machine, the Fast Data Finder made by TRW, as the clear leader for applications to pattern searching in biological sequences. This machine also has the potential for performing at least some aspects of the problem of determining sequence homologies. Over the past year, we have been negotiating with TRW to attempt to find some mechanism for BIONET's access to an FDF machine, but to no avail. We have been unable to meet their charges for access within our limited budget, although negotiations are continuing for access to less expensive versions of the FDF. Despite these problems, we continue to feel that the FDF is an extremely promising architecture, and we will do everything we can to arrive at an agreement with TRW.

In a parallel effort, we have entered into discussion with Dr. Peter Denning, head of the RIACS (Research Institute for Advanced Computer Science) project, headquartered at NASA/Ames Research Center. He

has a proposal submitted for purchase of a Connection Machine from Thinking Machines Corporation in Cambridge, MA. This is a processor that includes up to 64K individual processors operating in parallel. We have observed this machine in operation, performing high speed text searches, and its performance is impressive. BIONET has agreed with Dr. Denning to lead a Biotechnology Working Group, one of several groups that will explore a wide variety of applications of this novel architecture. At the time of writing of this report, final word on funding of the machine had not yet been received.

III.A.4. BIONET Training Program

Our training program during 1986-1987, emphasized the holding of regional trainings in different areas of the United States. The goal has been to try to draw BIONET users to trainings who would probably not otherwise attend. In addition to the shorter trainings at both the ASBC meeting in Washington, DC and the Miami Mid-Winter Symposia in Florida, we have held two magnet trainings during the summer of 1986: one in the northeastern United States at Dartmouth College in New Hampshire and one in the western United States at Stanford University in California. (See Appendix VI for an example of the mailer for BIONET training sessions). These trainings accomplished the main objective of training relatively novice users to more effectively utilize the system. Since these magnet trainings were successful, we plan to continue holding them in 1987.

In addition to the regional trainings we have also initiated a telephone training program for new users. The aim is to reach new users before they start using the system and assist them in getting started. Although we only began this program in August, the feedback that we have received so far indicates that it is worthwhile, and we plan to continue it.

The following summarizes the past year's training activites and those that are planned prior to the end of the current grant year. Training session schedules for all the trainings are listed in Appendix VII.

MIAMI MID-WINTER SYMPOSIA 1986 Training. BIONET sponsored a six hour training split between the afternoons of February 5-6, 1986 at the Miami Mid-Winter Symposia in Miami, Florida.

AMERICAN SOCIETY FOR BIOLOGICAL CHEMISTRY MEETING Training. We held a three hour class June 12, 1986 at the ASBC meeting in Washington, DC. The class was very well-attended (25 people) and we plan to hold future trainings at upcoming ASBC meetings.

DARTMOUTH Training. With much help from Bob Gross and Jo Steele at Dartmouth, BIONET held a two day training session on August 7th and 8th, 1986. There were 17 attendees, almost a third of whom (5) were from outside the area. The training was divided into novice (first day) and advanced (second day) with essentially all trainees attending both sessions. Dartmouth facilities included a VAX computer and terminal room for hands-on sessions for both days. The training was very well received.

STANFORD Training. We held a three day training for 43 people at the Stanford Business School's computer facility August 27-29. The three days were divided between novice, intermediate, and advanced topics. Although the training was rated highly by the attendees (80%), some felt that the number of trainees was too high. In the future we will try to limit class size to a more manageable 20-25 people so that more personal help can be given to the individual trainees.

MIAMI MID-WINTER SYMPOSIA 1987 Training. We plan to hold a training session on the evening of February 11, 1987 and it will be very similar to the previous year's session.

III.A.5. Resource Facilities

Previous reports have discussed the DEC-2060 and the various software and database libraries provided by the BIONET Resource. In this section we highlight significant changes and additions to the suite of hardware and software that comprise BIONET.

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

Hardware. The BIONET Central Resource Machine is a Digital Equipment Corporation 2060 computer. An Ethernet interface was added this year to provide access to ARPANET and other IntelliGenetics resources. The old DECNET front end, the DN20, was retired upon the introduction of the new Ethernet interface, the NI20. The MCA25 cache memory was ordered and scheduled for installation in December 1986, replacing the MCA20. This upgrade increases the throughput for the users of the resource.

The hardware configuration is as follows:

```
KL10-E Model R Processor:

      2 MF20/MG20 Memory controllers
   2 MW MG20 Memory
 .75 MW MF20 Memory
        MCA25 Cache Buffer Memory
      2 RH20 Massbus Channels
        NI20 Ethernet Interface

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

 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 1986.

Total disk space	433,000	(pages--222 million words)
Overhead/Common	<143,000>	(Core, System and System Support Libraries)
Swapping Space	< 25,000>	
File system Overhead	< 67,000>	(Directories and index pages)
	198,000	
BIONET Allocation	99,000	(Half of the available space)
BIONET Usage 12/86	< 85,000>	
Unused space	14,000	(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.

Public Data Network Connection. BIONET is accessed principally over the Telenet Public Data Network, operated by US Sprint. 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 Telenet 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.

Our former Public Data Network, UNINET, merged into Telenet in October 1986. BIONET had formerly made use of Telenet. UNINET was 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. The favorable pricing arrangement was renewed for a one year term beginning in July 1986, and continues to be honored by the new US Sprint Telenet.

During the year we increased from 12 to 16 the number of data network host ports 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 8 ports. The usage on these first 8 ports therefore represents many more

sessions than does the usage of ports 9 through 16. 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 9 ports fixed, 7 on traffic, and will change this distribution as required for the lowest possible cost.

We had been examining the replacement of the leased HPAD supplied by UNINET with a BIONET owned HPAD. The consideration is the savings of lease charges while maintaining adequate reliability. We were unable to procure a suitable HPAD for BIONET's use on Uninet. Following the Uninet merger into Telenet, we are reexamining the issue, and we may purchase an HPAD from Telenet rather than leasing it.

ARPANET - During the course of the year, BIONET participated in two task force meetings to look at issues related to computer networking in the scientific community. This task force operates in conjunction with the DARPA Internet Activities Board. BIONET has arranged Internet access to 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. The data connection to ARPANET took longer than expected, arriving in November 1986. The activation of the gateway computer at IntelliCorp is expected to take place during December. Since there are mail gateways from the ARPANET to other communications networks, this connection will do much to expand BIONET's reach, linking it with networks such as BITNET, EARN and CSNET as well as the DoD Internet.

Until the regular Internet connection is operational, BIONET has arranged mail access to the Internet with the valuable cooperation of the SUMEX-AIM NIH resource operated by Stanford. The same software that will link us to BIONET satellites allows us to exchange electronic mail through the SUMEX-AIM system. This has allowed scientists on BIONET to participate in information exchange using electronic mailing lists of Internet, which we import and make available as electronic bulletin boards on BIONET.

During the year, BIONET's central DEC-2060 resource was upgraded with the capabilities necessary for the Internet connection. An Ethernet interface now connects the DEC-20 to an IntelliGenetics local area network. An IntelliGenetics gateway connects that network to one at IntelliCorp, on which there will exist the gateway to ARPANET. Software for the TCP/IP protocols has been licensed and installed on the 2060.

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 29% of the machine, and Class II and staff 9%. The 24% overhead (system overhead, batch and computer staff and operations) is allocated one-half to BIONET, for a total of 50%. The only major change to these allocations since last year was an increase from 20% to 24% in the amount allocated to system overhead, reflecting effects of a new release of the operating system. Because the operating system itself tends to consume less overhead, the effects of the increase in other system overhead on the users are minimized.

Note that the BATCH class is assigned 1% of the system during prime time. In non-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 has been substantially greater than 50% of the total cpu cycles allocated. As an example, the percentage use of the machine for the month of October, 1986 is shown in Figure III-2. It is clear that BIONET is receiving far more than its allocated 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 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 utilized well over 50% of the total cpu cycles used on the 2060, and routinely consumes two-thirds 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-3 and III-4 by month, and totals. The cpu data in Table III-3 is also plotted in histogram form in Figure III-4.

The main conclusion derived from these data is that the BIONET resource is being fully utilized. The data for this year shows only a slight upwards trend and is on the average about 25% higher than the corresponding cpu usage for the previous year.

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

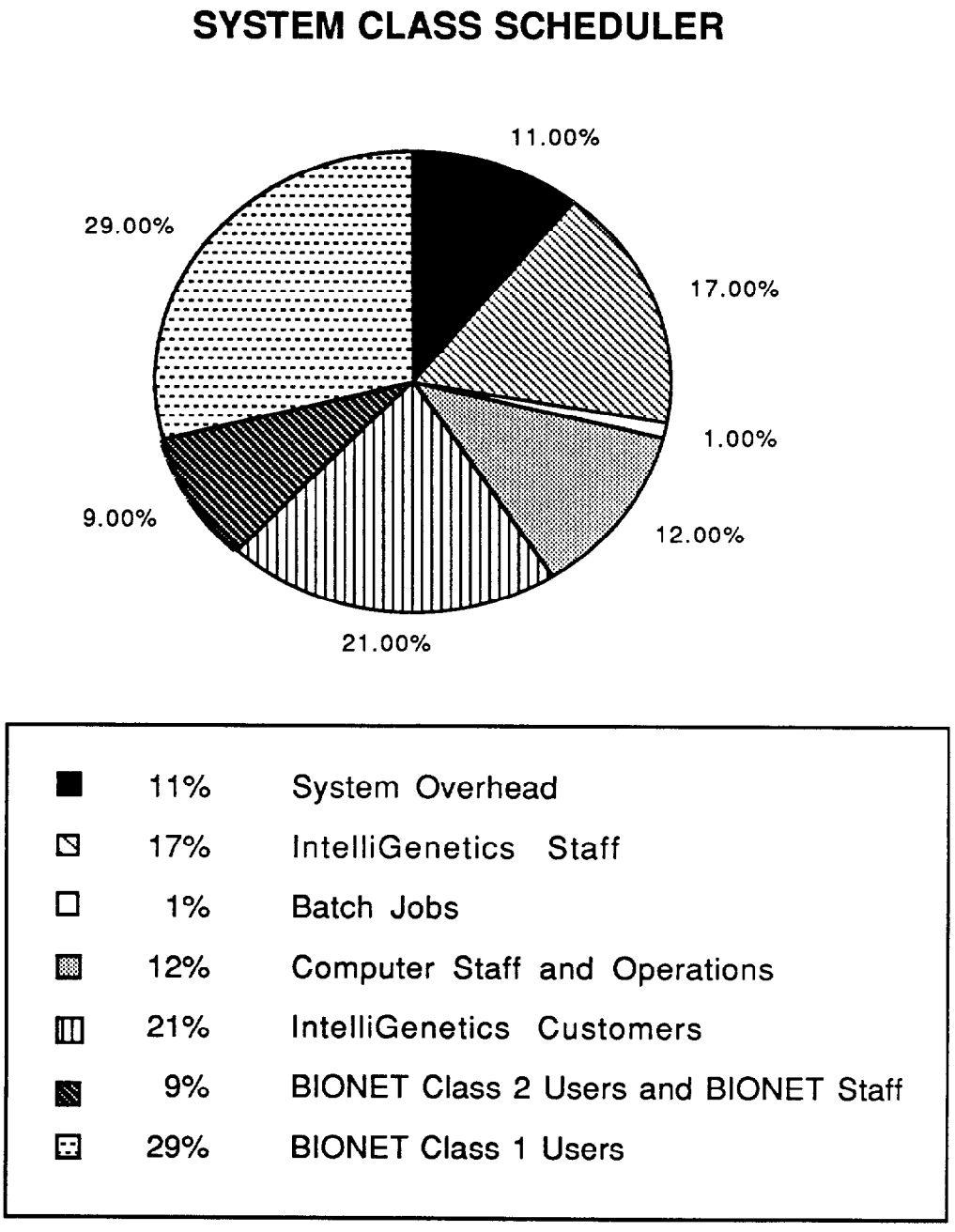


FIGURE III-2: Actual use of the DEC-2060 for the Month of October, 1986

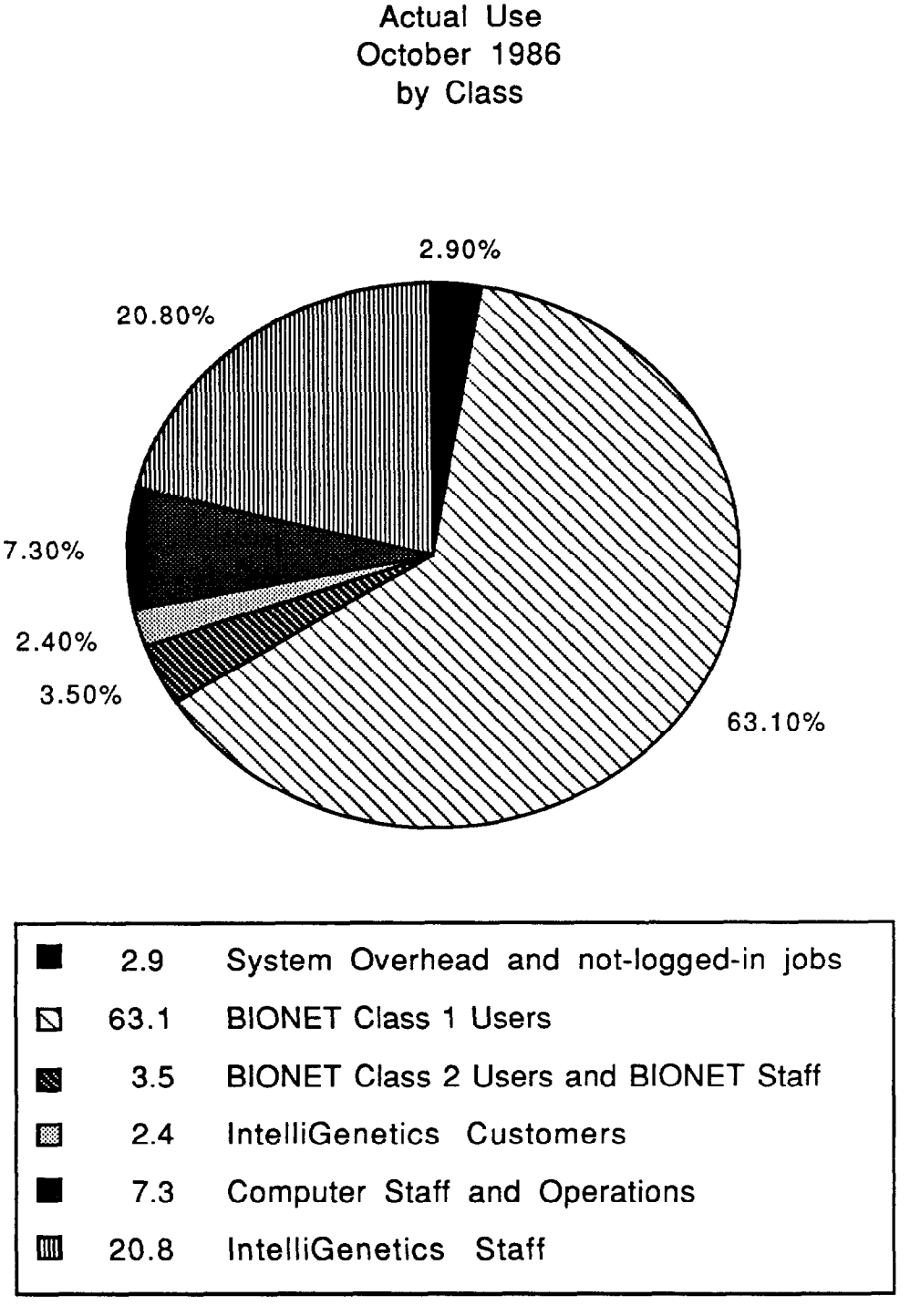
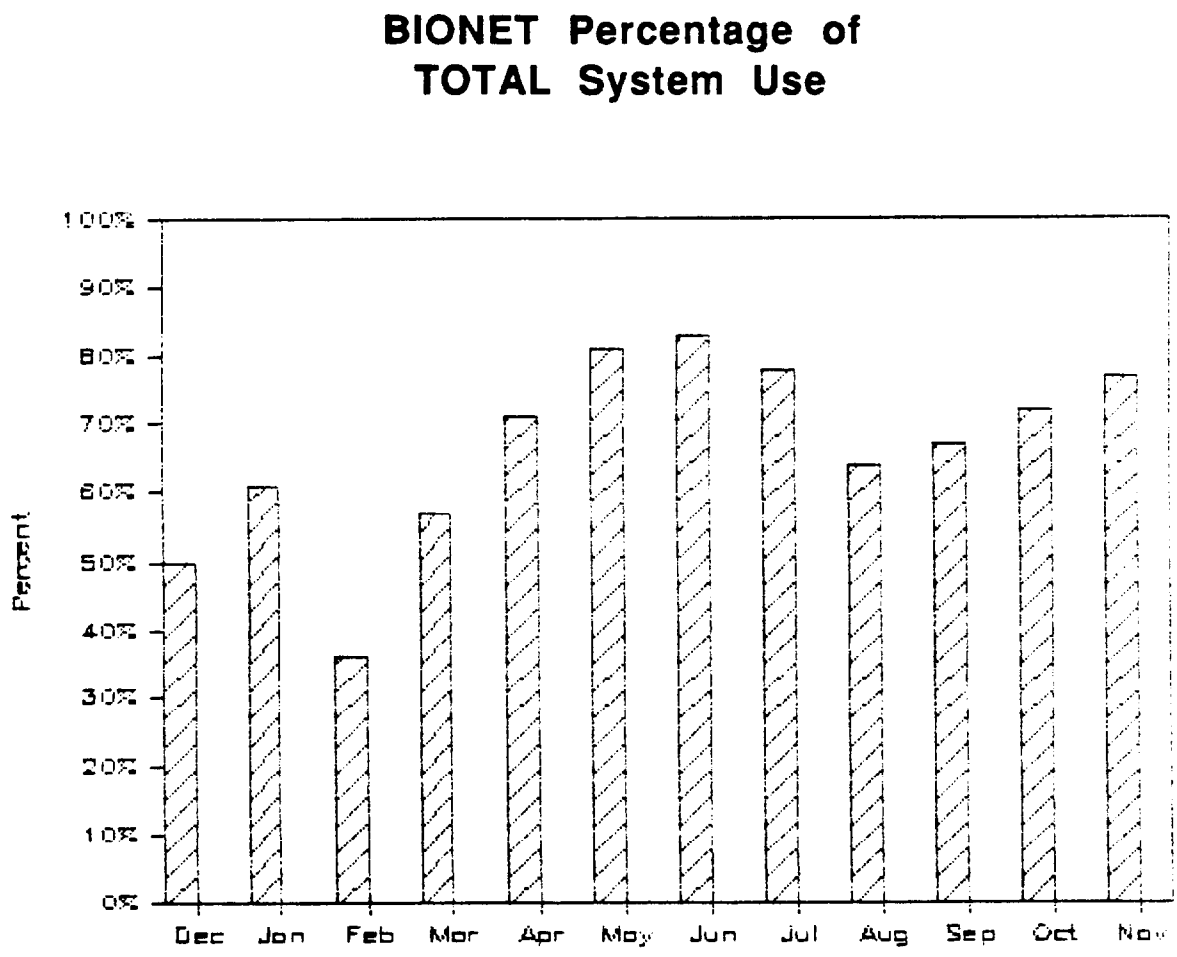


Figure III-3: BIONET's Percentage Use of the DEC-2060, 12/85 - 11/86



The total number of prime-time connect hours, (Table III-4), for BIONET Users is also up over last year by about 43%. The higher increase in connect time versus cpu time is probably a consequence of the increased demands on the system.

The data for non-prime time (weekends and 8 PM - 8 AM M-F) are shown in Tables III-5 and III-6, and the data on cpu time are plotted in histogram form in Figure III-5. Non-prime time cpu usage by BIONET has increased by around 75% over the course of the year. 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 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-7 and III-8 and the total cpu time is summarized in Figure III-6.

One important conclusion from all these data is that the Resource is close to 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.

Summary data for use of our telecommunications network are presented in Figure III-7 by month for the past 12 months' use of the UNINET (through the end of October, 1986) and Telenet (beginning October, 1986) networks. UNINET was taken over by Telenet during the course of this year, necessitating the change of networks.

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

There have been two major releases of the IntelliGenetics software systems that make up the Core Software Library. This software is made available to the BIONET community immediately upon its formal release. The first release, in mid-1986, included major upgrades to the heavily used nucleic acid (*SEQ*) and protein (*PEP*) analysis programs, and the sequence homology program *IFIND*, plus many improvements to other modules. For example, a DIGEST command was provided in *PEP* to simulate digestion of a protein with a selectable set of chemical and biochemical reagents, calculate fragment sizes, and predict gel behavior of the fragments.

The second release, around the end of 1986, includes substantial changes to all programs except *MAP* and *SIZER*. Major improvements to the *GEL* program for constructing a consensus sequence from individual sequences have been made, including the ability to combine sequencing projects, and to produce laboratory notebook records of all steps of the gel sequence assembly. In addition, an extensive graphical

Table III-3: BIONET Prime Time CPU Minutes

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	5732.4	409.5	360.6	6502.5
Jan	5293.4	425.4	337.7	6056.5
Feb	4406.8	234.8	268.4	4910.0
Mar	6181.9	384.2	346.8	6912.9
Apr	6303.0	721.5	386.2	7410.7
May	9274.3	668.9	305.7	10248.9
Jun	8014.1	389.7	301.3	8705.1
Jul	9171.4	245.0	216.2	9632.6
Aug	5857.3	615.4	430.1	6902.8
Sep	6083.8	834.2	321.0	7239.0
Oct	7815.1	601.7	398.2	8815.0
Nov	6161.5	418.8	493.7	7074.0
Total	80295.0	5949.1	4165.6	90409.7

Table III-4: BIONET Prime Time Connect Hours

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	2115.3	399.0	1943.2	4457.5
Jan	1967.8	680.3	1717.9	4366.0
Feb	1982.8	496.6	1488.8	3968.2
Mar	2364.5	633.2	1699.4	4697.1
Apr	2428.6	730.3	1702.7	4861.6
May	3515.5	674.6	1877.6	6067.7
Jun	2533.5	513.8	1642.0	4689.3
Jul	3213.8	679.9	1769.6	5663.3
Aug	2549.4	604.3	1785.3	4939.0
Sep	2442.6	834.5	1668.6	4945.7
Oct	3410.3	1049.5	2003.8	6463.6
Nov	2777.5	828.9	1563.3	5169.7
Total	31301.6	8124.9	20862.0	60288.5

Figure III-4: BIONET's Prime Time Use of the DEC-2060 12/85 - 11/86

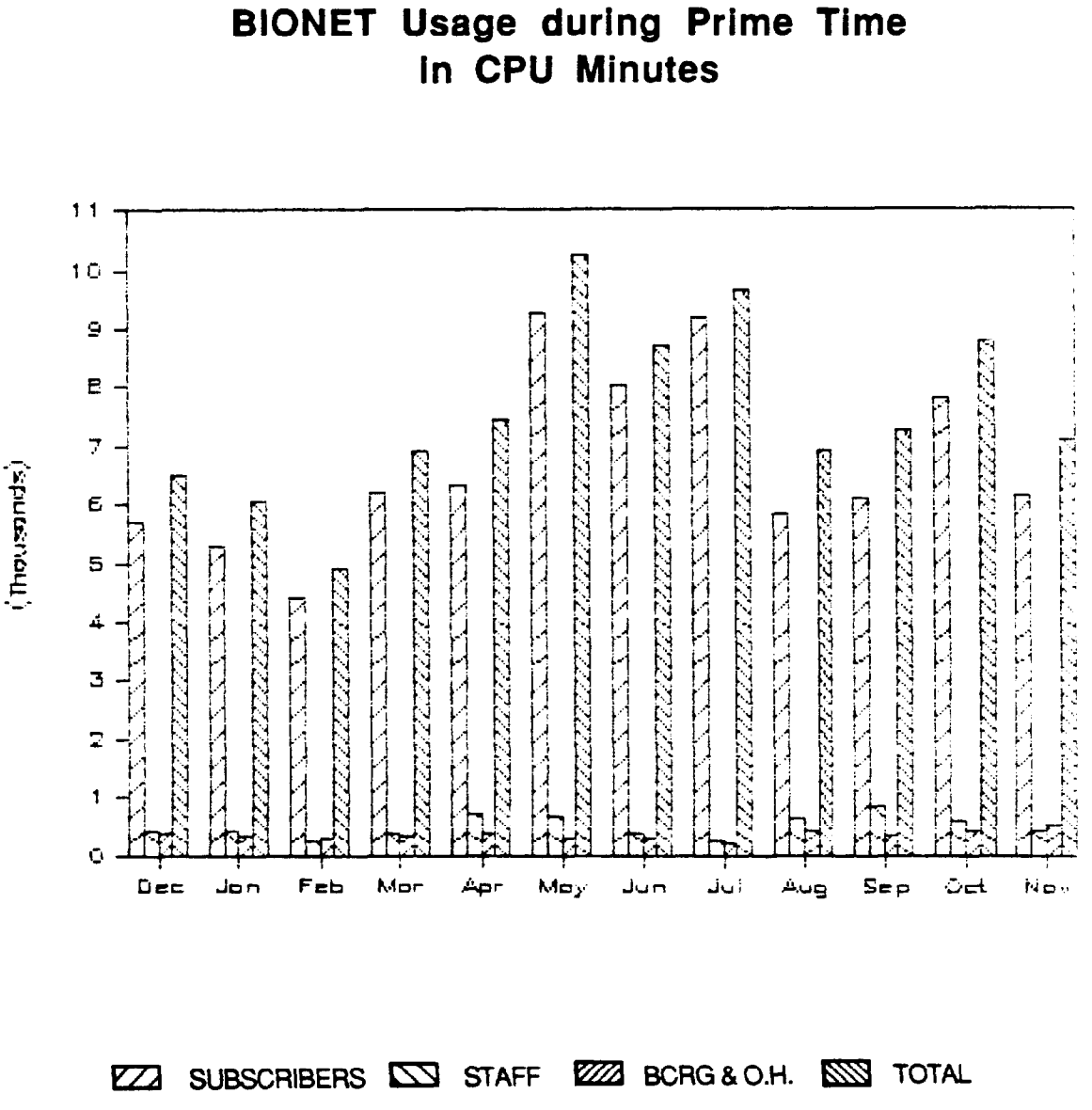


Table III-5: BIONET Non-Prime Time CPU Minutes

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	5888.9	81.7	132.4	6103.0
Jan	5258.4	51.3	940.8	6250.5
Feb	3398.4	31.6	930.5	4360.5
Mar	7836.9	131.2	1007.6	8975.7
Apr	6884.3	173.0	865.2	7922.5
May	15266.7	428.3	962.4	16657.4
Jun	10809.2	60.0	960.6	11829.8
Jul	8905.3	41.5	1038.7	9985.5
Aug	5847.1	46.6	1133.6	7027.3
Sep	7859.8	273.1	905.6	9038.5
Oct	8888.3	302.0	954.0	10144.3
Nov	9877.6	322.6	1107.0	11307.2
Total	96720.9	1942.9	10938.1	109601.9

Table III-6: BIONET Non-Prime Time Connect Hours

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	1311.0	56.3	305.3	1672.6
Jan	1230.1	144.2	2515.0	3889.3
Feb	1317.1	79.8	2362.3	3759.2
Mar	1698.3	113.2	2601.4	4412.9
Apr	1640.7	151.2	2595.7	4387.6
May	2559.7	161.4	2626.9	5348.0
Jun	1769.3	90.6	2362.8	4222.7
Jul	2011.4	196.0	2869.3	5076.7
Aug	1467.9	179.5	2472.9	4120.3
Sep	1416.8	174.3	2736.4	4327.5
Oct	2000.2	155.7	2612.4	4768.3
Nov	1863.0	134.8	2356.7	4354.5
Total	20285.5	1637.0	28416.8	50339.3

Figure III-5: BIONET's Non-Prime Time Use of the DEC-2060, 12/85 - 11/86

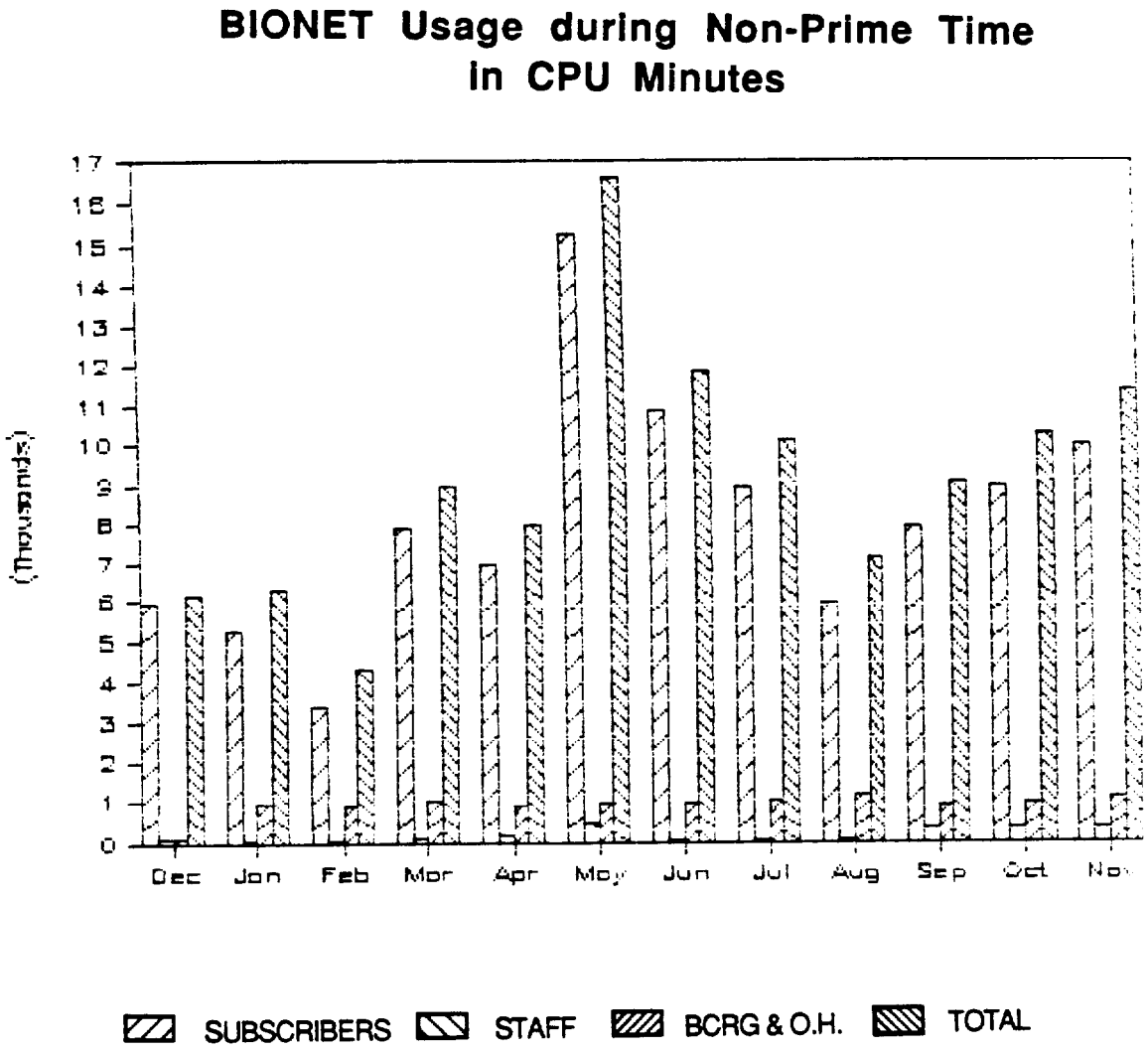


Table III-7: BIONET Total CPU Minutes

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	11621.3	491.2	493.0	12605.5
Jan	10551.8	476.7	1278.5	12307.0
Feb	7805.2	266.4	1198.8	9270.4
Mar	14018.8	515.4	1354.3	15888.5
Apr	13187.3	894.5	1251.4	15333.2
May	24541.0	1097.2	1268.1	26906.3
Jun	18823.3	449.7	1261.9	20534.9
Jul	18076.7	286.5	1254.9	19618.1
Aug	11704.4	662.0	1563.7	13930.1
Sep	13943.6	1107.3	1226.5	16277.4
Oct	16703.4	903.7	1352.2	18959.3
Nov	16039.1	741.4	1600.6	18381.1
Total	177015.9	7892.0	15103.7	200011.6

Table III-8: BIONET Total Connect Hours

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	3426.3	455.3	2248.5	6130.1
Jan	3197.9	824.5	4232.8	8255.2
Feb	3299.9	576.4	3851.1	7727.4
Mar	4062.8	746.4	4300.8	9110.0
Apr	4069.3	881.5	4298.4	9249.2
May	6075.2	836.0	4504.4	11415.6
Jun	4302.8	604.4	4004.8	8912.0
Jul	5225.2	875.9	4638.9	10740.0
Aug	4017.3	783.8	4258.1	9059.2
Sep	3859.4	1008.8	4405.0	9273.2
Oct	5410.5	1205.2	4616.2	11231.9
Nov	4640.5	963.7	3920.0	9524.2
Total	51587.1	9761.9	49278.7	110627.7

Figure III-6: BIONET's Total Use of the DEC-2060, 12/85 - 11/86

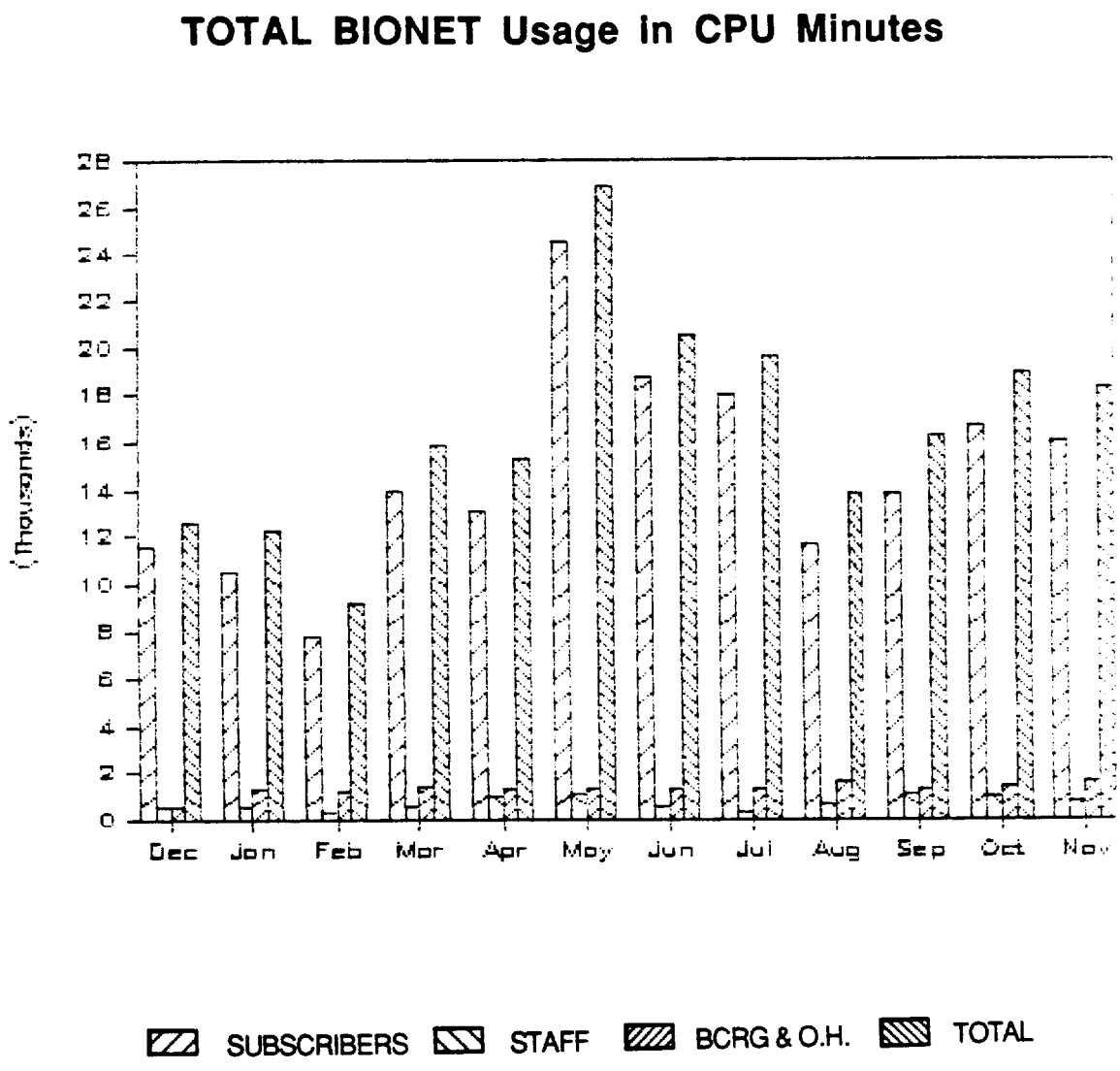
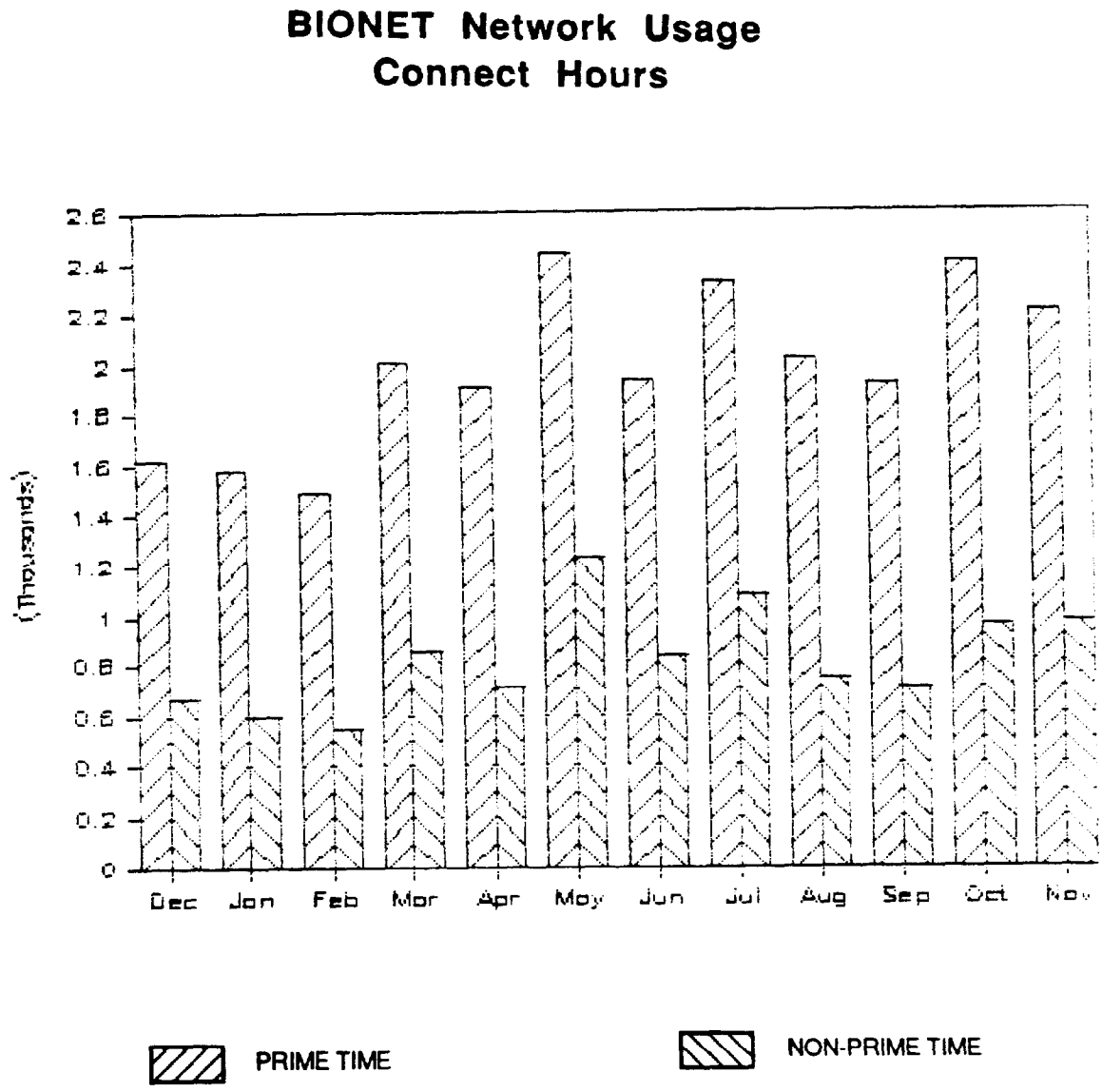


Figure III-7: Total Telenet and UNINET Network Use, 12/85-11/86



interface to the *CLONER* program will be available. This release will be followed shortly by the release of two new modules, *DDMATRIX* for sophisticated dot matrix comparisons of two sequences, and *GENALIGN*, for alignment of multiple nucleic acid or protein sequences.

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.

Operating System - A new release of the TOPS-20 Operating system was installed. This is version 6.1, the version currently supported by Digital Equipment Corp. Several new features are of use to BIONET, including the previously mentioned TCP/IP communication protocol (section 3.A.5.a). Other features include updated utilities, improved filename parsing, and directory access groups which can extend across file systems.

Programming Languages - Several updated versions of the C programming language from Stanford University and SRI International were installed during the course of the year.

System Utilities - A utility TFIND was created to enable users to find their TELENET network access phone number conveniently.

A utility to read Unix TAR format tapes was installed.

A utility MPW was created to suggest memorable, but suitable unique, passwords.

Communication - Several TCP/IP related utilities were installed. FTP permits file transfer under TCP/IP. TELNET permits virtual terminal connections under TCP/IP.

A new version of the BBOARD program permits full integration of electronic bulletin boards with the facilities of the MM message reading program.

III.A.5.e. Computer Software - Contributed Library

Software contributed to BIONET is placed in the <CONTRIBUTED> directory on the DEC-2060, to which only the BIONET community has access. Major software packages produced by BIONET collaborators and implemented on BIONET with the aid of our staff have been summarized under *Collaborative Research*section III.A.2.a, above.

III.A.5.f. Database Library

BIONET provides its users with a large number of different databases in support of molecular biology and molecular genetic research. These range from biological sequence and structure databases through bibliographic and genetic map databases. This year two different DNA sequence databases, the EMBL and NIH Genbank databases have grown in parallel and the overlap in their content has increased as they approach a more common format. Figures III-8 and III-9 show the growth in these databases with each release. In 1985 we maintained one release of these databases on BIONET for each release from the source. During the past year we have made one release every two months. In addition we have prepared special update files that allow the users to screen or search just the new sequences in each update. This has been a considerable simplification for the BIONET users and has helped save on CPU time as well. Our database releases have usually occured 2-3 weeks after obtaining the tapes from NIH GenBank or the EMBL.

The DNA sequence databases are used by BIONET scientists both as a source of sequence data and for homology searches. Two major types of searches are performed. Those involving the QUEST program search the DNA databases for interesting consensus sequences of functional importance. The second type of database search looks for similarities or homologies with sequences using the IFIND program. It is hard to estimate the total number of such searches since most users search different parts of the database rather than the entire database. We do know that the most frequently searched portions of the database are the files containing human DNA sequences. They are searched an average of 166 times per month (5-6 times per day). The rate of use of other sections of the database vary between 20 and 60 times per month. Perhaps a better measure of the use of the databases comes from the number of times that the QUEST and IFIND programs are used on the system. The QUEST program is used 864 times per month to search for consensus sequences and IFIND is used 1115 times per month to search for homologies. This amounts to nearly 2000 searches utilizing the databases each month.

The EMBL database is used considerably less than the NIH Genbank database with its files being referenced only 27 times per month. We attribute this to the great amount of overlap between the NIH GenBank and the EMBL databases as well as to the more frequent releases from Genbank. As these two resources work more closely with each other, BIONET may chose to release only the NIH Genbank database.

Protein Sequence Database - BIONET makes the Protein Identification Resources protein sequence collection available for similar homology searching. This database is searched 468 times per month for homology using the IFIND program and the same database is searched 60 times per month using Bill Pearson's XFASTP program. This makes protein database searching the second most highly used database on the BIONET resource. The growth of this database and the number of releases are shown in Figure III-10.

Restriction Enzyme Databases - Thanks to Dr. Richard Roberts, one of the members of the BIONET National Advisory Committee, we have established one of the most up-to-date lists of restriction enzymes available. Dr. Roberts maintains a restriction enzyme registry and distributes his updated lists on a monthly basis in an electronic message to BIONET. BIONET brings these new lists into use within about 24 hours. These lists are used within the core programs SEQ and PEP and are referenced 592 times per month. Recently Dr. Robert's group has included with each update a description of the new enzymes and new data about existing enzymes. This material is available on-line.

Vectorbank - Vectorbank is a collection of restriction maps of important cloning vectors, viruses and phages that is maintained by the IntelliGenetics staff. This database is used by the CLONER program for manipulating restriction maps and simulating DNA cloning experiments. Initially many of the restriction maps in Vectorbank were taken directly from sequence information in the sequence databases. The current database contains many restriction maps contributed directly by BIONET users or by scientists that have developed the vectors for various uses (sequencing vectors, expression vectors etc.).

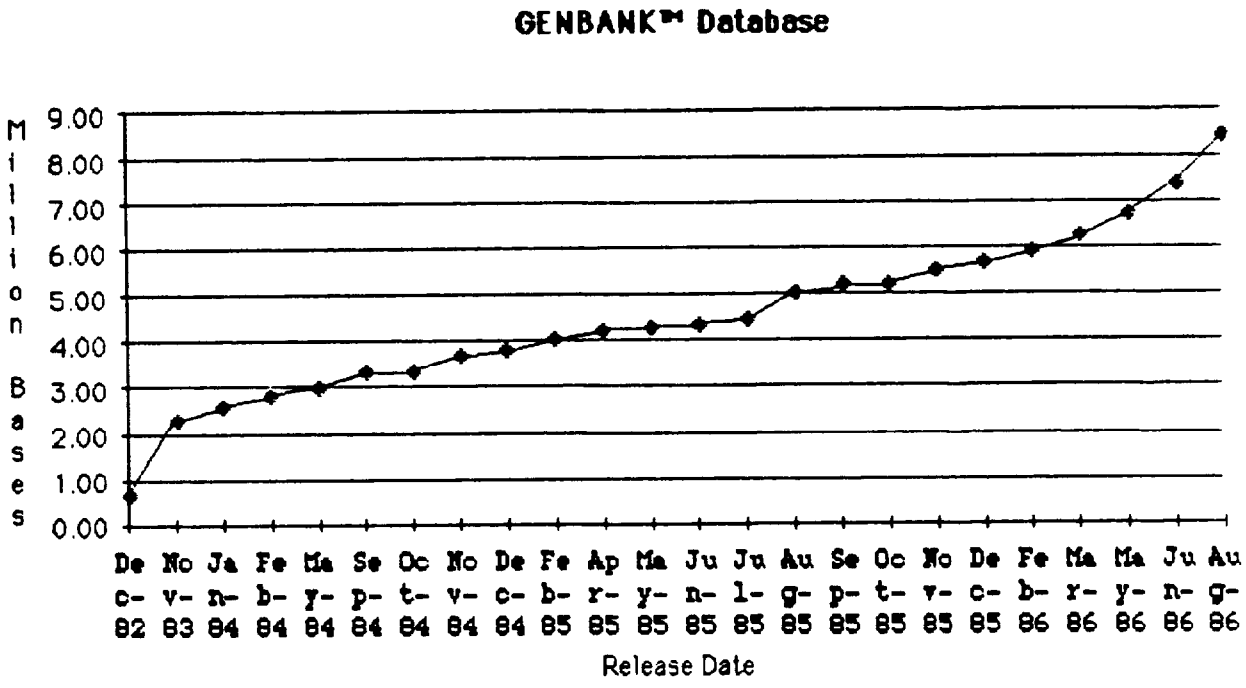
We also contacted Elsevier who have recently published the book *Cloning Vectors*, containing restriction maps of over a thousand vectors and phages, but unfortunately the information in that book is not available in a computer readable form. Nevertheless, the book provides an important resource and as individual scientists enter the map data into the computer, either through the Cloner or Strategene programs, we can make them available to the entire BIONET community.

The vectorbank is referenced about 70 times per month and the CLONER program itself is used 131 times per month.

Consensus Sequence Oligonucleotide Databank - Recently Drs. Trifonov and Brendel have published a compendium called *Gnomic* of over 800 consensus sequences from the literature. This book contains a list of short oligonucleotide sequences which are either tandem repeats, promoter sequences, operators, protein binding sites, restriction enzyme cleavage sites, or other short sequences of functional importance. Dr. Trifonov has already made this list available to us and has requested his publisher to make the entire book available in a computer readable form. We intend to edit this list of sequences into a set of KEYs that can be used for database searching via the BIONET QUEST program. We will call this database a "KEYBANK". In addition to the sequences themselves we intend to include the literature references that reported the sequence and a cross reference to the GNOMIC book. The QUEST program is already used for this purpose but the collection of KEY's that we have is quite small and is collected in a non-systematic fashion. QUEST is currently among the most heavily used programs with 864 references per month and with an extensive databank of keys we expect that usage to increase.

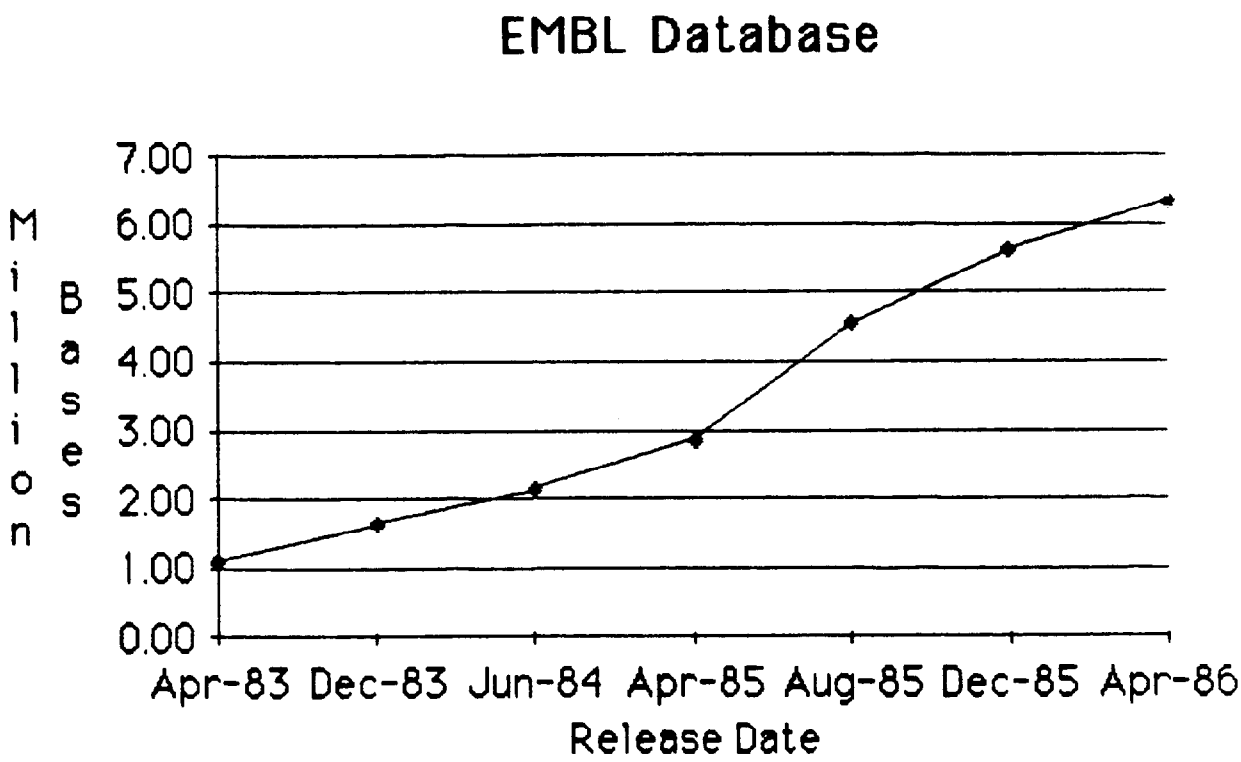
Genetic Map Of *Drosophila melanogaster* - Last year we made a computer readable version of Lindsley

Figure III-8: Size and Release Dates of the NIH GenBank Database



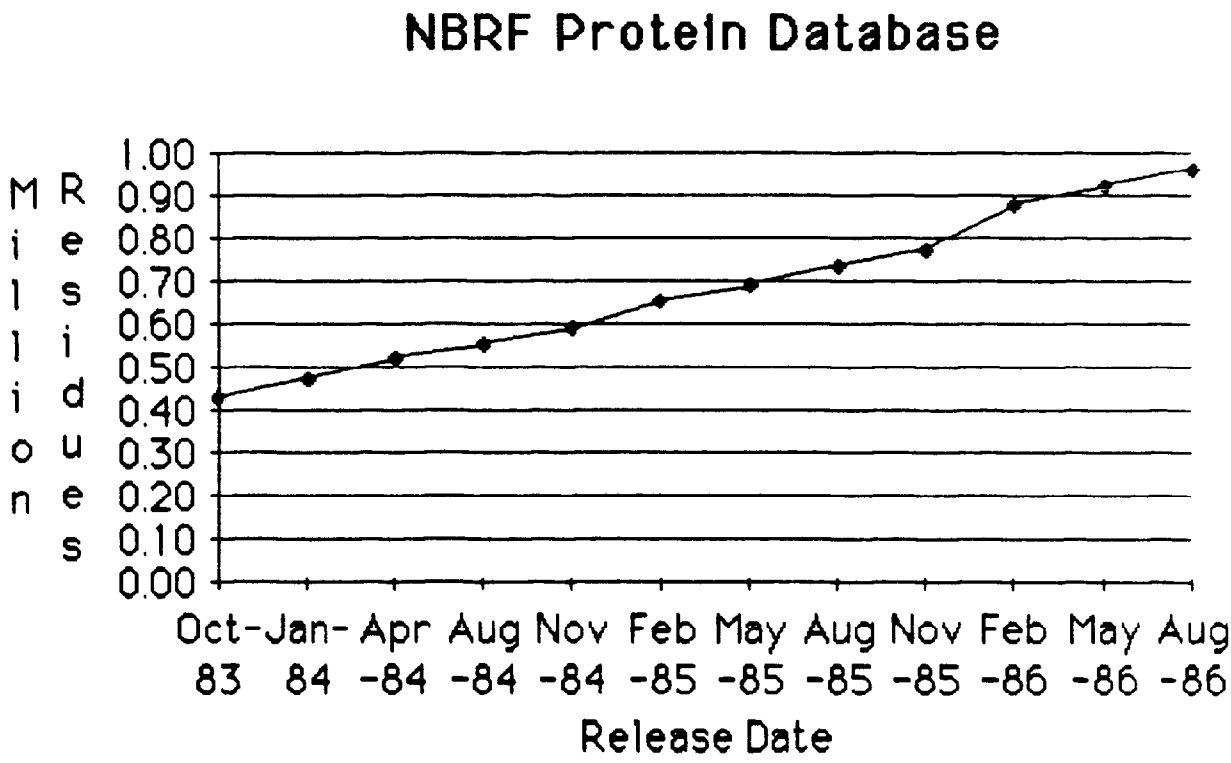
The size and release dates of the NIH GenBank database on the BIONET computer. This figure includes a few releases prior to the BIONET grant.

Figure III-9: Size and Release Dates of the EMBL DNA Sequence Database



The size and release dates of the EMBL DNA sequence database

Figure III-10: Size and Release Dates of the PIR Protein Database



The size and release dates of the Protein Identification Resource protein database maintained by the National Biomedical Research Foundation.

and Grell's *Genetic Variations of Drosophila melanogaster* available on line. We described how to search this text for genetic markers and rearrangements, and, most importantly, we described how to find all genetics markers within a specified region. Despite the fact that only 15 BIONET scientists work with *Drosophila*, this database was referenced 79 times for a rate of about 10 references per month.

Dr. Lindsley is currently updating this 18 year old reference work which has more than trebled in size. During this year we have mounted two updates to this work and now have the entire list of mutants as of 1986 available for searching. In addition we are working with Dr. John Merriam at UCLA who is preparing a molecular correspondence between cloned gene segments and the genetic map. Dr. Merriam's work is a database of all mapped cloned sequences and provides an important reference for starting cloning experiments in this organism.

Drosophila is a very good organism for presenting the entire genetic map on BIONET because the data is completely available in a computer readable form. As more and more genetic maps become available we will consider the best way to provide them to the BIONET community. Dr. Frank Ruddle has also made the genetic map of humans available as a Spires database at Yale. We will attempt further collaborations where this seems useful and feasible.

Brookhaven Protein Structure Database - Although we have no programs for displaying three dimensional protein structure on the BIONET computer, there are a number of BIONET users that have access to both mainframe and more recently, microcomputer programs for displaying and editing this type of information. Several of our users have requested that we make this type of information available. We have obtained the database from the Brookhaven Laboratory and mounted its files on BIONET. While the entire database is too large for BIONET users to download into their microcomputer, it should be a valuable resource for those wishing to obtain a single protein structure or those wanting the most recent updates to their own library. We are considering other ways of making this three dimensional data of value to the BIONET community.

III.B. Highlights

Total usage on the system has increased by 25% over the last year. About 170 inactive accounts were replaced by 130 active new users.

We have taken several steps to improve the electronic communications available through BIONET, including substantial work on revamping the bulletin board system and installation of mail forwarding among facilities via Telenet and ARPANET.

We have begun the collection and dissemination of several programs designed to help solve the very

important problem of alignment of multiple biological sequences. This important Core Research activity will place BIONET in a leadership role in making such programs routinely available.

A new system of on-line help menus was installed to assist scientists in using the resource.

The Brookhaven Protein Databank is now available on BIONET along with a new facility designed to expedite dissemination of sequence data.

III.C. Administrative Changes

There have been many administrative changes within BIONET during the past year. These have come about for reasons ranging from a change in ownership of IntelliGenetics itself, to personnel shifts, additions, and resignations. We are fortunate that none of these changes has had a significant impact on the Resource itself, in particular, its appearance and availability to the community. Here is a brief summary of changes:

- In early May, 1986, IntelliGenetics became a Joint Venture between Amoco and IntelliCorp, 60% owned by Amoco and 40% by IntelliCorp. As part of this transition, the management of the Cooperative Agreement was turned over completely to IntelliGenetics, and Dr. Dennis H. Smith replaced Dr. Ralph Kromer as Principal Investigator.

- In late April, 1986, Dr. Marcia Allen took a leave of absence from her position as a BIONET Scientific Consultant; this position was filled in early August, 1986 by Dr. David Kristofferson. Dr. Kristofferson has recently been appointed Acting Resource Manager.

- Mr. David Horner joined the Computer Facilities Group as a Senior Operator in early June, 1986, replacing Ms. Mary Yardley, who transferred to another position in IntelliGenetics. Mr. Horner spends 50% of his time on BIONET-related tasks concerning the operation of the DEC-2060.

- On July 1, 1986, Ms. Nancy Bigham, a BIONET Scientific Consultant, transferred to IntelliGenetics Customer Support, concurrent with the transfer of Ms. Theresa Friedemann from Customer Support to BIONET Scientific Consultant.

- Mr. Rob Liebschutz joined the Computer Facilities Group as a systems programmer in early November, and is spending about 50% of his time with BIONET working on BIONET Satellite and ARPANET communications.

- Dr. Sunil Maulik joined IntelliGenetics in early November, 1986, and is working 100% time as a BIONET Scientific Consultant.

- Dr. Dennis Smith resigned from IntelliGenetics on December 1, 1986 to pursue other opportunities. His position as Principal Investigator has been taken by Dr. Michael Kelly, President of IntelliGenetics.

The resignation of Dr. Smith is the only administrative change that might temporarily affect the operation of BIONET. However, as part of his transition, arrangements have been made for commitment

of additional time of the BIONET Co-Investigators Drs. Brutlag, Friedland and Kedes. This will ensure a smooth transition while a search for additional senior staff for BIONET takes place.

III.D. Resource Advisory Committee and Allocation of Resources

The membership of BIONET's National Advisory Committee remains unchanged from last year. It consists of:

- Professor Joshua Lederberg, MD, PhD. (Chair), President, The Rockefeller University.
- Dr. Saul Amarel, PhD., Director, Information Systems Technologies Office, Defense Advanced Research Projects Agency, Department of Defense.
- Professor Alan Maxam, PhD., Dana Farber Cancer Institute, Harvard Medical School, Harvard University.
- Dr. Richard J. Roberts, PhD., Senior Staff Investigator, Molecular Biology, Cold Spring Harbor Laboratory
- Thomas Rindfleisch, MS, Director, Knowledge Systems Laboratory, Department of Computer Science, Stanford University.
- Professor Charles Yanofsky, PhD., Department of Biological Sciences, Stanford University.
- Professor Fotis Kafatos, PhD., Department of Cellular and Developmental Biology, Harvard University.

Meetings of the Committee are normally held once a year. The last regularly scheduled meeting occurred at the end of the previous grant year, on February 24, 1986 in Mountain View. At this meeting we reviewed the progress of the Resource and discussed directions for Core and Collaborative Research for the current grant year covered by this Report. In particular, we discussed our goals for pursuing three Core Research projects, multiple sequence alignment, network communications associated with the BIONET Satellite Program, and special hardware for rapid text searching. All three projects were given the strong endorsement of the Committee; progress in the three areas was summarized previously under *Core Research*, section III.A.3, above.

Another strong recommendation of the Committee was that BIONET should be truly international in scope, and that efforts to extend communications and collaborations to other resources in foreign countries should be encouraged. We have been hampered in carrying out this recommendation because of delays in the installation of ARPANET, but this will be completed shortly. When installed, we will immediately implement electronic mail facilities and routing directories so that communication among international sites is made simple. Several such sites were mentioned previously under *Collaborative Research*,section III.A.3 above.

The Committee agrees with our methods for allocating the Resource. The DEC-2060 computer uses its windfall scheduler to allocate cpu time to the various categories of users and overhead, as described in our previous Report. The cpu time is distributed on a first-come, first-served basis. This method has been very successful, with considerably more than BIONET's 50% time allocation being delivered to BIONET scientists (see section III.A.5.b, above).

We continue to request that the community not have more than one person per PI group using BIONET at the same time during prime time. The community continues to do an excellent job in complying with this policy.

We continue to allocate additional disk space to PI groups involved in managing large sequencing projects or extensive databases of sequences. We do this on an *ad hoc* basis upon requests by investigators. We project that sufficient disk space will be available through the remainder of this grant year, and probably beyond, to meet the storage requirements of the community. Our archive and retrieval system is working smoothly to store seldom-used files on tape and retrieve them promptly (1-2 days) upon requested.

III.E. Dissemination of Information of Resource's Capabilities

We discuss two areas related to dissemination of information about the Resource that we have pursued this grant year. The first is interactions with the scientific community through participation at conferences. The second is use of the electronic mail and bulletin board facilities of the Resource itself to keep the BIONET community aware of changes and improvements.

III.E.1. Community Interactions and Awareness

We have used two methods this year to inform the community about BIONET and to solicit applications for access to the Resource. The first method has been participation at major conferences where we have presented papers and/or have had booths at exhibitions. These efforts are summarized previously under Training, Subsection III.A.4. At these conferences, we have distributed the standard applications packets to scientists, after demonstrating to them the capabilities of the Resource. (See Appendix VIII for an example of the renewal newsletter).

Last year we placed one advertisement in *Nucleic Acids Research* and while we did receive some response, most of the queries came from Soviet bloc countries. We received a greater response from our presence at the above mentioned conferences and focussed our advertising efforts in that direction. See Appendix IX for a copy of the advertisement which appeared in the Miami Mid-Winter Symposia, 1987 brochure.

This year was the first year that we solicited the renewal of subscription fees. We used this opportunity

to bring our subscribers up to date on the changes to the resource. A sample of the newsletter is in Appendix X.

III.E.2. Electronic Communications

The electronic communication facilities of BIONET provide another important way to disseminate information about the Resource. In addition, electronic mail and bulletin boards provide a mechanism for scientific and technical interchanges among members of the community.

III.E.2.a. Bulletin Boards

The electronic bulletin boards are an important component of the BIONET Resource. They provide BIONET users with a facility for the exchange of data, laboratory techniques and ideas. Our users represent a wealth of knowledge. Communication is the key to accessing and disseminating that knowledge.

Information on the current status of the Bulletin Boards has been previously discussed in Section III.A.2.e.An announcement used to attract new bulletin board leaders is given in Appendix XI.

III.F. Suggestions and Comments

In the previous year's Annual Report, we made two strong suggestions for improvement. The first was a plea for more warning about the effects of Federal budgetary decisions on our grant award. We are pleased that we had ample warning of potential cuts this year, which dramatically improved our own planning processes and ability to react. We are not pleased that cuts had to take place, but we did appreciate the warning.

Our second suggestion was related to the lack of NIH initiatives in computer networking, making it difficult for computer resources to share information electronically. We have discussed this problem at length with the BRTP staff and with our counterparts at other resources. The net result is general agreement that: 1) there are already many efforts aimed at coordinating the growth of computer networks; 2) a new effort may be counterproductive; and 3) the NIH in general, and BRTP in particular, should do all it can to encourage its grantees to participate in existing networks. We feel the best way to achieve (3) is to encourage persons submitting proposals to anticipate the costs of networking, and to ensure that Study Sections and Councils understand the importance of maintaining these funds in final awards.

I. BIONET RESOURCES REAPPLICATION

9 December 1986

Dear BIONET Principal Investigator:

The National Institute of Health requires, as part of our Annual Report, that we review the status of all BIONET subscribers each year. Thus, we need information from you on any changes, from your original application, with respect to institutional affiliation, address, funding status, sub-investigators, etc. Most importantly, we need a list of all publications in which BIONET played a role; as well as three copies of the reprint or preprint. We also ask that at this time you reaffirm your original agreement for access to BIONET.

We are enclosing a printout of the record we now have on your lab. Please indicate on the printout any changes in:

- Your title, affiliation, mailing address and/or phone number
- The list of your sub-investigators who have a subdirectory in your account.

On page 1 of the BIONET Resource Reapplication please type or print your full name and title. Read and affix the date, name of official and signatures to the BIONET agreement.

On page 2, please type or print your name and note change in status of funding and provide a list of current publications resulting, in part, from the use of the BIONET Resource (Remember to cite the BIONET Grant # 1 U41 RR-01685-03 in all such publications.)

ON page 3, type or print your name and provide a brief description on how BIONET was used in your research.

We would also like to give you this opportunity to comment on the BIONET resource - what role it is playing in your research and any suggestions/requests for improvement.

Because we must prepare our Annual Report in December, we need you to return this re-application to us no later than November 7, 1986. Thank you for your cooperation.

Sincerely,

Mary Lou Warner
BIONET Administrator

BIONET RESOURCE

Reapplication

Fiscal 1987

Principal Investigator (full name and title):

BIONET Agreement

As Principal Investigator of this grant to use the BIONET Resources, I agree to adhere to all conditions and restrictions for use of the BIONET Resource, as described in the document "The BIONET^{tm} Resource, Description and Applications Form" and such further regulations as may be issued from time to time by the NIH or BIONET's National Advisory Committee.

The BIONET Resource will not be used for any commercial purpose which is not specifically identified to and approved by the NAC. Any pertinent change in sponsorship, continuity of grant support, or use made of BIONET will be reported promptly to the BIONET Resource Manager.

I have also furnished a copy of this re-application to the Grants Administrator of my institution, whose signature appears below.

I also assume full responsibility for all users listed on this applications form and will monitor their compliance to the conditions and restrictions for access to the BIONET Resource. I will inform the BIONET Consultant, (electronic mail address: BIONET), by electronic mail, immediately about any changes in this group of users, i.e., departure of existing user or addition of new staff qualified to use the resource. I will inform new users of the above mentioned conditions and restrictions.

Date:

Name of official:_____

Signature of Principal Investigator

Signature of Grants Administrator

BIONET re-application page 2

Funding Status

Please note any new funding, including Institution, Grant Number, title of grant, and duration of grant.

Current Publications

List current publications resulting, in part, from the use of the BIONET Resource (use standard bibliographic format). Remember to cite the BIONET Grant # 1 U41 RR-01685-03 in all such publications. A sample citation would be: *Computer resources used to carry out our studies were provided by the N.I.H. sponsored BIONET*^{tm} *National Computer Resource for Molecular Biology.* **Please submit 3 copies of preprints or reprints.** If you would like to have any of your other publications listed on the BIONET Bibliographies Bulletin Board, please list those on a separate sheet.

PI Name:_____

BIONET Re-application Page 3

Use of BIONET

Briefly describe how BIONET has helped your research:

COMMENTS

We invite your comments, suggestions and requests about the BIONET Resource. Which programs are the most useful to you - the least? Should the bulletin boards be broader in scope - more specific? Would you like more interaction with other users? What else would you like to see included in the BIONET Resource, for example, other computer programs. Would you like more information about the BIONET on-site package?

II. ON-LINE HELP MENUS

@HELP ME

WELCOME !!!!
This menu is designed to provide answers for common questions.
Please send any comments about it by electronic mail to BIONET.

FOR INFORMATION ON:	TYPE AFTER THE @
Support Phone Numbers for BIONET users	HELP CONSULTANT
Support Phone Numbers for Commercial Users	HELP SAR
Making Printed Copies of Information and Files	HELP HARDCOPY
How to Use the Editors	HELP ED
How to Find Other Users on the System	HELP WHOIS
How to Use the Electronic Mail	HELP E-MAIL
How to Use Bulletin Boards (and SAVE $400!!)	HELP BULLETINS
Scientific Meetings by Computer !!	HELP MEETINGS
A Guide to the Main IntelliGenetics Programs	HELP MAIN-PROGRAMS
Contributed Programs and FREE PC Software	HELP SOFTWARE
File Copying, Deleting, etc.; CTRL Key Use	HELP TOPS20
Transferring Files Between Computers	HELP FILE-TRANSFER
Nucleic Acid, Protein, and other Databases	HELP DATABASES
How to Find Sequence Files	HELP FIND-FILES
Running Time-Consuming Programs Via Batch Jobs	HELP BATJOB
TELENET & Phone Connection Problems	HELP TELENET

To display this menu again: HELP ME

@HELP FIND-FILES

NOTE: Detailed information on sequence file names for the various databases can be found on pages 24-29 in the Introduction to BIONET (BIONET users) or pages 32-38 in the IntelliGenetics Timesharing system manual (commercial users). Please note the following revision to the section on Protein Sequences in those manuals. The file <IG>SNBRFPEP.LIST has been changed to <IG>SNBRF.LST in conformance with the naming conventions for the other analogous database files.

FOR INFORMATION ON:	TYPE AFTER THE @
How to use and interpret *'s in filenames	HELP FF-WILD
Finding a file that contains the sequence of a specific gene or protein	HELP FF-FIND
Choosing which files to search in the IFIND and QUEST programs using:	
the GenBank database (for nucleic acids)	HELP FF-XNA
the EMBL database (for nucleic acids)	HELP FF-XNA
the NBRF database (for proteins)	HELP FF-PRO
To display the main menu again:	HELP ME

III. XMULTAN ALIGNMENT

```
CONSENSUS       PLPAATL TYCANTTTT TTGC JAAALNTGTTTT

1                GTAAATA TCCACTTTT TTGC AGAGTCTGTTTT
2                GTAAATA TCAACTTTT TGGC AAAATCCGTTTT
3                  GAAAT TTCATTTTT TTGC CAAAAGTATTTT
4                AAAAATATTTTAAATTT TGCC AAAATCCGTTTT
5                  GAAAT TTCATTTTT TTGC CAAAAATATTTT
6                GAAAATA TCCGATTTT TTACAAAAATTTTTTTT
7               GATGATTT TTGGATTTTGT  C CGAAAATGGATT


C TYC APATTTTGPTCPTMAAAAAAATAATCAGT TTTTTKPYCAMAACAT

1 TCC AAATTTCGGTCATC    AAATAATCATT TATTTTGCCACAACAT
2 TCC AAATTTCGGTCATC    AAATAATCAGTGTTTTCTGCTACAACTT
3 CCC AGATTTT  TTGTGAAAAAAATATTTGGT ATTTTGATCAAAACAT
4 TTC AAATTTTGATCAT    AAAATAATCAGT TTTTTTGCAACAACTC
5 CCC AGATTTT  TTGTGAAAAAAATATTTGGT TTTTTGATCAAAACAT
6 TTCGATTTTTTTGATC     AAAAATAATCCGT TTTTTTTAAATAACCA
7 TTC TGTTTTTTTGCGATTTAAA   TATCAGT ATTATGATCAGAACAT


C TJPAAATAATTGTCYPAATATGGAATGTCATACCTCG TTGAGTTCGTAA

1 AAAAAATAATTGTCTGAATATGGAATGTCATATCTCA CTGAGCTCGTAA
2 TAAAAACAATTGTCTGAATATGGAAACTCATACGTCG CTGAGCTCGTAA
3 TCGAAATAATTACCCAAATATGGAATGTCATACCTCGTTTGAGTTTGTTT
4 TAAAAATAAATGTCTGAATTCGGAATGGTATACCTCG TTGCGTTCGTAG
5 TCGAAATAATTACCCAAATATGGAATGTCATACCTCG TTGAGTTTGTTT
6 CAAAAATAGTTGTCCAAA GTGGAATGCCATACCTCG TTGAATTCGTAA
7 TCGAAATATTGGTCCAAATATGGAATGTCATATCTCG TTGAATTCGTAA


C YTLAATTYCCAATCGAACTGTGTTCAA   AAMTTGGAAATTMLA TTTK

1 TAAAATTTCCAATCAAACTGTGTTCAA   AAA TGGAAATTAAA TTTT
2 TTAAATTTCCAATCAAACTGTGTTCAA   AAA TGGAAATTAAA TTTC
3 CTTAAATCCCAATCGAATTGCGTTCAA   GTTTTGG AATTCTA GGAG
4 TTTAATTTCCAATCGAACTGTGTTTAA   AAGTTGGAACTCTAT TTTT
5 CTTAAATCCCAATCGAATTGCATTGCATTCAAGTTTTGAATTCTA GGAG
6 CAAAATTCCCTATCGATCCATGTAC A   TACTTTGAAATTCTA ATTG
7 TTAAATTTCCAATCGAACTGTGTTTAC   CAA  AAAAAATGAATTTTT
```

```
C TTTKCJATTTTTTGCAAATTTTGATGATGNTACCCCTTACAAAAAATGCG

1 TTGGCCACATTTTGCAAATTTTGATGACCCCCCTCCTTACAAAAAATGCG
2 TTTGACATAGTGTGCAAATTTTGATGATG       TTACAAAATATGTG
3 GTTTCAGTTTTTTGCAAATATTGATGA   TACCCCTTACAAAAAATTCG
4 TTTGCCATTTTTTGGCAAATTTTAATGATGTTACCCCTTACAAAAAATGCG
5 GTTTCAATTTTTTCCAAATTTTGATGATGGTACCCCTTACAAAAAATTCG
6 TTTGCCATTTTTTCGCAAATTTTGATGATGGTACCCCTTATCGAAAATGCG
7 TTTTTAATTTTTTTCCAATTTTTTGAT

C AAAATTTGNCCAAAAATTAATTTNNCLAAATCJKTNAAAA   AGTGATA

1 AAAATTGATCCAAAAATTAATTTCCCTAAATCCTTCAAAA   AGTAATA
2 AAAATTTGCCAAAAAATTGATTTCTCTAAATCCTTGAAAA   AGTAATA
3 AAAATTTGGCCAAAAATTAATTTTACAAAATCAGTTTAAA   AGTGAAA
4 AAAATCGACCCAAAAACGAATTTCCC AAATCCGTCAAAA   AGTGATA
5 AAAATTTGGCCAAAAATTAATTTTACAAAATCAGTTTAAA   AGTGAAA
6 AAAATTTGTCAAATTTTTTTTTTGCGAAAATC   GAAAAAGTAGGGATA

C GGGATNGTTAGCANTGGTLATTAGCLGCLCAAAACAGTNNTTCTTTYAKC

1 GGGATCGTTAGCACTGGTAATTAGCTGCTCAAAACAGATATTCGTACATC
2 GGGATCGTCAGCACTGGTAATTAACTGCTCAAAACAGTTTTTCATGCATC
3 GGGGTTGTTAGTATTGGTTGTAAGGAGTACAAAATGGTACTTCTTTT GC
4 GGGATCGTTAGCATTGGTAATTAGCTGCTCAAAACAGTTATTCTTTCATG
5 GGGGTTGTTAGTATTGGTTGTAAGGAGTACAAAATGGTACTCCTTTTTGC
6 GACATAGTTAGCTATCTTTATTAGCAGCACAAAACAGTCTTTATTTTAGC

C TNTNTGACCATTTTTAGCCAAGTTATPPCNAAAA

1 TATGTGACCATTTTTAGCCAAGTTATAACGAAAATTTCGTTT
2 TATATGACCCTTTTTAGCCAAGTTATGACAAAAATTTCGTTT
3 TCTCTGACCATTTTTAGTCAAGTTATAGCCAAAACAGCCAATTT
4 TTTATGATCATTTTTAGCCAAGTTATGATTAAAATGCCAATA
5 TCTCTGACCATTTTTAGTCAAGTTATAGCCAAAAAAGCCAATTT
6 TGTGCGTCCATTTTTAACCAAGTTATGGCCAAAACGCCTATT
```

APPENDIX : Best alignment of seven Drosophila Satellite DNS's as determined by XMULTAN on the BIONET resource. The satellite sequences are from the 1.688 g/cm3 satellite DNA cloned from 1) D. melanogaster, 359 bp repeat, 2) D. melanogaster, 353 bp repeat, 3) D. mauritiana, 4) D. orena, 5) D. simulans, 6, D. yakuba, 7) D. teissieri.

IV. PROGRAMS CONTAINED ON THE IBM AND MAC DIRECTORIES

ARC.DOC	Documentation for the ARCHIVE program. This program can pack many files into *.ARC files and also extract files from these compacted forms. This aides in downloading multiple file packages. This file is also contained in the ARC51.COM file below.
ARC51.COM	Down load this binary file and then run ARC51. This program unpacks itself resulting in the generation of ARC.EXE and ARC.DOC version 5.1. Then delete ARC51.COM. ARC.EXE can then be used to unpack *.ARC archive files.
ARCE.ARC	ARCE is a much smaller version of ARC that can extract files from *.ARC archives but cannot pack files. It is useful if you only download *.ARC and never need to prepare *.ARC archives.
BINHEX.BAS	BINHEX is a 189 line basic program that can convert binary files to hex files and back again for up and downloading over seven bit data paths. The program HC (hex convert) is much faster and is preferred.
BINHEX.HLP	Help documentation for BINHEX.BAS.
EMACS.ARC	Archive file containing a micro computer version of EMACS. Although only a limited subset of EMACS commands are available, this program can edit several files simultaneously and can edit very large files.
EMACS.DOC	Documentation for microcomputer EMACS. Also contained in EMACS.ARC.
HC.ARC	Archive file containing HC.COM and HC.DOC. See below.
HC.COM	Binary file for HC program that converts *.EXE and *.COM programs into hexadecimal HEX files or converts HEX files back into *.EXE or *.COM files. Needed to transfer binary files over seven bit data paths or networks.
HC.HEX	A HEX version of HC.COM. You must have BINHEX.BAS or a HEXCONV to convert this file back to HC.COM after downloading.
KERMIT-V120.EXE	An old compact version of KERMIT that will run on all versions of MS DOS and is relatively bug free. IT is particularly good for use on floppy disks where space is limiting. It is missing many features of MSKERMIT but it an excellent terminal emulator.
LUE210.COM	A program for extracting files in the compressed *.LBR

library format. Rename this file to LUE.COM after you have downloaded it.

LUE210.DOC — Documentation for LUE120.COM

LUU208.LBR — The complete LUU.COM program and documentation. The LUU Library Utilities allow you to pack and squeeze files into *.LBR format for up and downloading. This program is useful if you intend to upload *.LBR files or maintain them yourself. If you merely want to download and extract files from libraries, then use LUE210.COM instead.

MODEM7.ASM — Part of the assembler code for the parameter section of the public domain MODEM7 program. Assembler programmers can use this file together with DEBUG to alter the default speed and port settings for MODEM7.

MODEM7.COM — A public domain version of MODEM that allows multiple file transfers to and from most other XMODEM programs. The program is set up to send data out COM1 port at 1200 BAUD but these parameters can be changed by program commands or by altering the program itself with the aid of the MODEM7.ASM file above. The terminal emulator of MODEM7 is very poor and setting TERMINAL ADM3 is recommended.

MODEM7.DOC — Manual for use of the MODEM7 program.

NUSQ110.COM — New version of the Unsqueeze program that can unsqueeze *.?Q? files. This program is needed if any of your downloaded files are in the squeezed format. Rename this program to NUSQ.COM or USQ.COM after downloading.

NUSQ110.DQC — Documentation for NUSQ110.COM. You must convert this *.DQC file to a *.DOC file by running NUSQ110.COM before attempting to read it.

SQ129.COM — Program to squeeze files. Useful to compact files before uploading them to minimize communication time and/ or disk space. The program converts *.COM files to *.CQM files, *.DOC files to *.DQC files, *.exe files to *.EQE files etc for many file types. Any file with a Q in the second position of the extension should be assumed to be in squeezed format and should be unsqueezed before using.

SQ129.DQC — Squeezed form of the documentation for SQ129.COM

STRIP.LBR — Program and documentation for converting Wordstar files to standard ASCII files. It strips out the eighth bits that Wordstar leaves and adds normal carriage returns and line feeds. This is necessary before attempting to upload wordstar files to Bionet or before sharing the

files with other word processors. STRIP can also convert TABS in ASCII files to spaces and also convert files with lots of spaces to TABS.

The following programs are available on the MAC: directory:

BINHEX4.BAS This Basic program when run will produce a Binhex 4.0 application on your disk. This application will convert hex files (*.HEX) files, compressed hex files (*.HCX) and squeezed hex files (*.HQX) to normal Macintosh files.

BINHEX4.DQC A squeezed form of the Binhex4 manual. Run Binhex4 on this document to generate a MacWrite file of the manual.

BINHEX4.HQX A Binhex4 compressed HEX file of the Binhex4 application.

BINHEX4.PAS A pascal program which when run will generate the Binhex4 application. After running BINHEX4.PAS you may delete it keeping only the resulting application.

EMACS.HQX A microcomputer version of the mainframe EMACS program that runs on the Macintosh. The file must be downloaded and then converted with Binhex4 to generate the application.

FREETERM18.HQX A public domain terminal emulator and XMODEM file transfer program. This program is a poor terminal emulator (TERMINAL ADM3) but it does a good job of up and downloading files using the MODEM file transfer program. Convert the file FREETERM18.HQX to the Freeterm application using Binhex4.

MAKE-MAKERS.HQX This program converts an existing application on the Macintosh into two files. The resulting *.PAS file is an Apple Pascal program which when run will regenerate the desired application. The resulting *.BAS file is a Microsoft Basic program which when run will regenerate the application. This program was used to generate the BINHEX4.BAS and BINHEX4.PAS files listed above. The MAKE-MAKERS.HQX file must be converted with Binhex4.

PACKIT.DQC A Binhex4 version of the Packit Manual. Convert with Binhex4 to obtain a Macwrite document. The file must be converted with Binhex4 to generate the PackIt manual.

PACKIT.HQX Packit is a program that packs and unpacks several files into one single file for up and downloading purposes. It is similar to the Library and Archive programs for the IBMPC. PACKIT takes several files including applications and documents and puts them into a single

*.PIT file. After files are packed by PACKIT then they are usually converted to a *.HQX file by Binhex4 before transmission. IF one downloads a *.HQX file and then converts it with Binhex4 and it becomes a *.PIT application, then PACKIT must be used to separate out the individual files before use. The file must be converted with Binhex4 to generate the PackIt application.

PACKITII.HQX Version II of the PACKIT program does more sophisticated data compression on the packed files. It is compatible with files packed with version I of PACKIT. The file must be converted with Binhex4 to generate the PackIt II application.

The following programs are available on the <PC-SOFTWARE.ZUKER> directory:

README.DOC Documentation on this software.

PCFOLD.EXE Packed version of PCFOLD.
PCFOLD2.EXE Version without packing of the arrays.

MENUDAT2 Data files needed by the program.
MENUDAT
MSG HLP

FOLD ENR Energy file.

FOLD BAT Batch file to run the program. This batch file suppresses the 8087 coprocessor before the call to PCFOLD.EXE.

FOLD2 BAT Batch file to run the program. This batch file suppresses the 8087 coprocessor before the call to PCFOLD2.EXE.

PSTV Default sequence file.

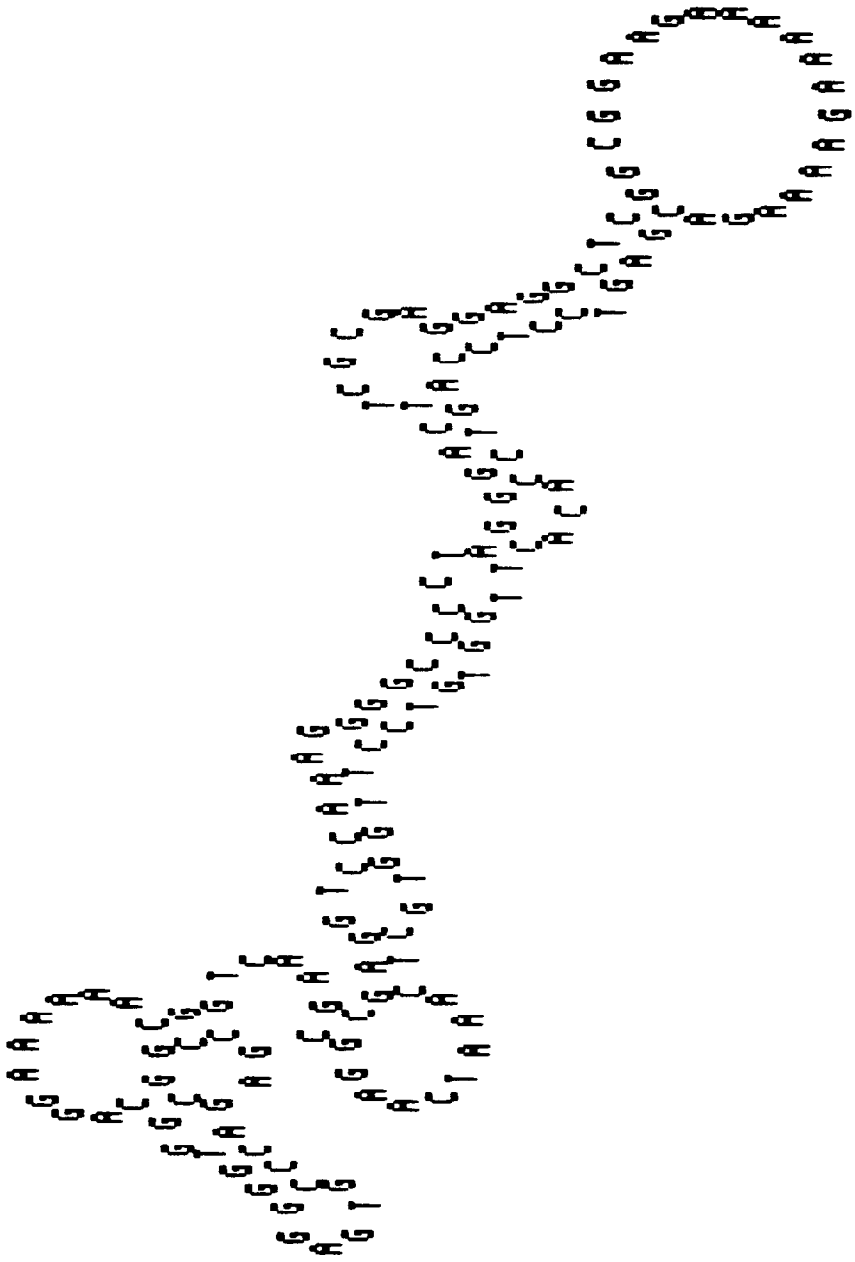
The directory <PC-SOFTWARE.THOMPSON> contains the following programs:

IMOLECUL.ARC Programs to display RNA structures graphically
HMOLECUL.ARC using *.CT files produced by Zuker's PCFOLD.
IMOLECUL requires IBM Graphics Card and
HMOLECUL requires Hercules graphics card.
Files must extracted with ARC or ARCE.

MOLECULE.DOC Manual for use of MOLECULE program.

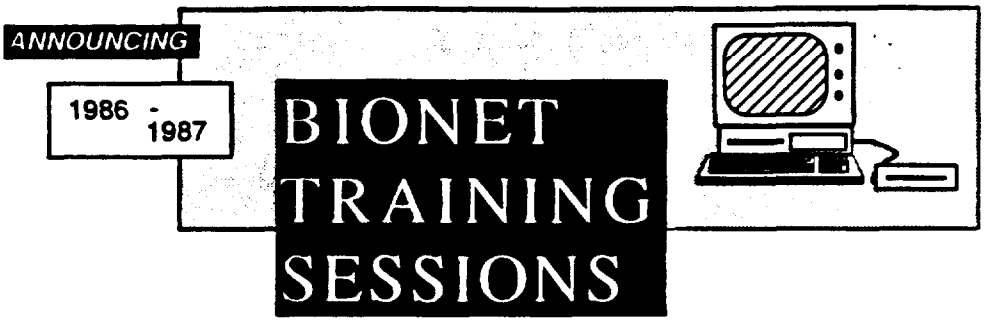
MOLECULE.PAS	Turbo Pascal source code for MOLECULE program. MOLECULE requires Turbo Graphix Toolkit to be compiled.

V. DISPLAY BY THOMPSON'S MOLECULE PROGRAM



APPENDIX : RNA SECONDARY STRUCTURE PREDICTED BY PF-FOLD and displayed by J.R. Thompson's molecule program. This figure shows the type of presentation possible with software contributed to the PC-SOFTWARE library. It shows the minimal energy RNA structure of a 150 base pair segment of a viroid RNA predicted by Zuker's PC-FOLD program and displayed by Thompson's MOLECULE program.

VI. MAILER FOR BIONET TRAINING SESSIONS



IN RESPONSE TO YOUR INTEREST . . . The BIONET consultants are arranging training sessions in your area. The following training sessions should be of interest to all BIONET users as well as others doing research in the field of Molecular Biology.

ASBC JUNE 12, 1986 WASHINGTON, D.C. 1:30 TO 4:30 P.M.

COURSE CONTENT

OVERVIEW OF BIONET AND THE COMPUTER SYSTEM

Core Program Library • Sequence Databases and Development Library • Electronic Mail and Bulletin Boards • System Commands and Directory Organization

SEQUENCE DATABASE ORGANIZATION AND SEARCHING METHODS

Rapid Sequence Alignment and Similarity Searches • Sequence Retrieval Using Exact or Ambiguous Patterns

OTHER CLASSES: THE FOLLOWING WORKSHOPS HAVE BEEN TENTATIVELY SCHEDULED. CONTENT WILL BE DETERMINED BY YOUR RESPONSE. PLEASE RETURN YOUR REGISTRATION ASAP.

DARTHMOUTH COLLEGE, HANOVER, N.H.	AUGUST 7 - 8
STANFORD UNIVERSITY, PALO ALTO, CA	AUGUST 27, 28, 29
UNIVERSITY OF IOWA, IOWA CITY, IA	EARLY SEPTEMBER
MIAMI MID-WINTER SYMPOSIUM	FEBRUARY 11, 1987 (Evening)

Emphasis will be placed on the hand-on use of the Intelligenetics programs as tools for research in Molecular Biology. Depending upon your response, the sessions will be conducted at novice and/or advanced levels. Tuition fees may be necessary to cover costs.

RETURN THIS PORTION TO BIONET C/O INTELLIGENETICS 1975 EL CAMINO REAL MOUNTAIN VIEW, CA 94040

Yes, I would like to take advantage of the BIONET workshops. Please send me confirming dates and fee schedule.

I am interested in ☐ NOVICE ☐ ADVANCED

I am interested in attending the following session

☐ DARTHMOUTH COLLEGE, HANOVER, N.H.
☐ STANFORD UNIVERSITY, PALO ALTO, CA
☐ UNIVERSITY OF IOWA, IOWA CITY, IA
☐ MIAMI MID-WINTER SYMPOSIUM

☐ I would be interested in hosting a session at my institution

Principle Investigator:_____
Sub-Investigator:_____
Institution:_____
Address:_____
_____
City_____ST____Zip_____

PLEASE CHECK THE TOPICS THAT ARE OF INTEREST TO YOU:

☐ Managing large DNA sequencing projects
☐ Restriction mapping tools
☐ Simulation and design of recombinant DNA experiments
☐ DNA or protein sequence database organization and searching methods
☐ Distributed processing: moving data to and from the BIONET computer
☐ Remote job processing - running programs away from the terminal
☐ DNA or protein sequence analysis
☐ Sequence comparision methods
☐ Use of BIONET on-line editors

VII. TRAINING SCHEDULES

BIONET™ TRAINING SCHEDULE

FOR THE

MIAMI MID-WINTER SYMPOSIUM

BIONET will be conducting three hands-on training sessions at the Miami Mid-Winter Symposium. BIONET subscribers are encouraged to attend the training sessions to become more familiar with the resource and thereby utilize it more fully. Investigators who are not subscribers are invited to learn how BIONET can aid in their research. The schedule for the training classes is as follows:

Wednesday, February 5 -- Orange Blossom Room

1:00 - 2:50 BIONET OVERVIEW
Programs and features that are available to BIONET users, how to access BIONET, how to send electronic mail and read or not read bulletin boards, how to manage the files in your BIONET account.

ENTERING SEQUENCES
How to enter and edit nucleic acid and protein sequence data.

3:10 - 5:00 DATABASE SEARCHING
How to search the NIH and NBRF databases for DNA or protein sequences.

Friday, February 7 -- Orange Blossom Room

2:00 - 4:00 RESTRICTION ENZYME MAPPING
How to generate restriction enzyme maps and create and use an individualized restriction enzyme list.

SEQUENCE ASSEMBLY
How to assemble gel sequence information to generate a consensus sequence.

In order to help cover the expenses of the training a small charge will be made. Cost of the courses is $30 for one, $60 for two and $80 for three. Please enroll at the IntelliGenetics booth #22 or at the Orange Blossom Room.

The BIONET computing resource is a cooperative agreement between IntelliGenetics and the National Institutes of Health. BIONET users have access to electronic mail and bulletin boards to connect them to other members of the research community. IntelliGenetics software available on BIONET allows users to gather, organize, store, and analyze DNA, RNA and protein sequence data.

5 U41 RR01685-03

BIONET™ Training Syllabus

ASBC

Thurday, June 12, 1986

1:30 - 4:30 pm

BIONET OVERVIEW

Programs and features that are available to BIONET users

- -Overview of the core programs
- -Databases on line
- -Contributed software
  - XFASTP
  - XFASTN
  - IDEAS
  - BIOFLD
- -Communications facilities
  - e-mail
  - bboards

Accessing BIONET

- -Using UNINET to connect to BIONET
- -Logging in procedures
- -Solutions to some common problems associated with log-in
  - UNINET error codes
  - Connected okay, but message is garbled
  - Backspace doesn't work

Sending electronic mail and using the bulletin boards

- -Basics of using the mail program
  - Sending mail /editing messages
  - Reading mail
  - Deleting old messages
- -Basics of the BBOARDS
  - Getting a listing of the Bboards
  - Accessing a particular bboard
  - Reading messages
  - Returning to your mail file

File management

- -Viewing files in your account
  - Use of wildcards with the DIR command

-Removing old files

-Archival and retrieval of files

-Important files to keep and files that can be removed

-How to see your disk space and disk usage

-Reallocation of disk space

SEQUENCE ENTRY

Entering and editing nucleic acid or protein sequences

-Use of the GENED program to enter sequences

EDIT line editor

ESEQ display editor

DATABASE SEARCHES

Locating a sequence or group of sequences in the nucleic acid or protein sequence databases.

-Use of the FIND program for rapid sequence location

Files to use with FIND

NIH.LST

EMBL.LST

SNBRF.LST

-Use of the QUEST program for sequence retrieval

When to use QUEST

Collection of similar sequences to a file

Searching for patterns within sequences

Rapid sequence alignment and similarity searches

-Using IFIND to align your sequence with similar sequences

Initiating a search with IFIND

READing in a .FND file and using it to align sequences

Dartmouth College

Bionet Training Schedule - Novice

August 7, 1986

9:00-10:15	Introduction - Overview of Bionet System: Login, System Commands, Mail, File structure, Databases
10:00-10:30	Break
	Overview of Programs:
10:30-10:45	GENED - Sequence Data Entry and Editing GEL - Sequencing Gel Management Program
10:45-11:15	SEQ - DNA Sequence Analysis Program PEP - Protein Sequence Analysis Program
11:15-11:35	SIZER/MAP - Restriction Enzyme Fragment Sizing and Mapping CLONER - Recombinant DNA Simulation System
11:35-12:00	QUEST - Database Similarity Searching IFIND - Database Similarity Searching

Novice Training Continued

12:00-1:00	Lunch

Hands on session:

1:00-2:15	GENED and GEL Programs
2:15-3:15	SEQ and PEP Programs
3:15-3:30	Break
3:30-4:15	SIZER and MAP Programs
4:15-5:00	QUEST and IFIND Programs

Dartmouth College

BIONET Training Schedule - Advanced

August 8, 1986

9:00-10:30	GEL - Searching and elimating vector sequences; Semi-automatic vs. automatic merging
10:30-10:45	Break
10:45-12:00	CLONER - Simulation of the construction of pUC9
12:00-1:00	Lunch

Hands on session:

1:00-2:30	PEP - Comparison of the Search and Align algorithms for protein sequence homology searching; setting chemical similarity matching for homology searches
2:30-3:15	QUEST - Searching using complex keys
3:15-3:30	Break
3:30-5:00	IFIND - Similarity searching between a translated portion of DNA and a QUEST retrieved portion of the NBRF database

Stanford University

Bionet Training Schedule

August 27, 1986

The morning session will be geared to the novice user:

9:00-10:15	Introduction Overview of Bionet System: Logging on, System Commands, Mail, BBoards
10:15-10:30	Break
10:30-12:00	Directories, File Structure and Location; Xsearch and Find
12:00-1:00	Lunch
1:00-2:15	GENED - Sequence entry and use of ESEQ editor
2:15-3:30	SEQ - Restriction enzymes site searching
3:30-3:45	Break
3:45-5:00	PEP - Designing probes with PEP; Hybrid protein construction and hydropathicity analysis

Stanford University

BIONET Training Schedule cont'd

August 28, 1986

9:00-10:30	Sequence Alignment Algorithms
10:30-10:45	Break
10:45-12:00	Database Searches (QUEST/IFIND)
12:00-1:00	Lunch
1:00-2:30	GEL - Sequencing Gel Management Program
2:30-3:45	SIZER/MAP - Restriction Enzyme Fragment Sizing and Mapping
3:45-4:00	Break
4:00-5:00	CLONER - DNA Cloning Simulation

Stanford University

Bionet Training Schedule - cont'd

August 29, 1986

9:00-12:00	Advanced Topics Including: QUEST Searching using complex keys IFIND similarity searching using a QUEST retrieved portion of a database.
12:00-1:00	Lunch
1:00-3:00	Editors, Batch Jobs
3:00-5:00	File transfer; up and downloading of files and programs between PC's and Bionet

VIII. BIONET APPLICATION

BIONET™

Dear Researcher:

You are invited to apply for access to the BIONET [tm] National Computer Resource for molecular biology. Enclosed are a description of BIONET, an application form, and order form for BIONET documentation.

The BIONET Resource is a central computer facility serving the computational needs, for both research and communication, of the molecular biology community. The Resource is funded by a five year, cooperative agreement with the Biomedical Research Technology Program, Division of Research Resources, National Institutes of Health. IntelliGenetics[tm], Inc. of Mountain View, California will provide the computer facilities, core software, and support. Responsibility for overseeing the Resource rests with a National Advisory Committee (NAC), comprised of Drs. Joshua Lederberg (Chair, Rockefeller), Saul Amarel (Rutgers), Fotis Kafatos (Harvard), Allan Maxam (Harvard Medical School), Thomas Rindfleisch (Stanford), Richard Roberts (Cold Spring Harbor), and Charles Yanofsky (Stanford).

The BIONET Resource has three goals:

- To provide computational assistance in data analysis and problem solving for molecular biologists and researchers in related fields.
- To serve as a focus for development and sharing of new software tools.
- To promote collaboration and rapid sharing of information among a national community of scientists.

Please read the enclosed User Agreement closely. By signing it, you will be agreeing to adhere to both the letter and the spirit of the guidelines described.

Each principal investigator must complete an application to be eligible to use the BIONET Resource. Access cannot be passed on from one principal investigator to another. Each scientist who qualifies for and currently has his or her own source of funding is considered a principal investigator.

Please type the information on your application form for legibility and accurate processing. Processing time will take approximately four weeks after receipt of your application.

If you are applying from a commercial or foreign organization, be sure that your application contains sufficient supporting material to allow the National Advisory Committee to make its judgements.

If your application is approved, we will send you a welcome notification, the "Introduction to BIONET" documentation, and instructions for logging on the BIONET

computer via the UNINET telecommunications network. We will also provide initial on-line training at your convenience.

Communications is a critical component of the BIONET Resource. On approval of your application, we will send you information on using Electronic Mail, Electronic Bulletin Boards, and File Transfer programs. These features will allow you to exchange information and ideas instantly with the BIONET staff and other users.

An annual fee of $400 is currently being charged to all U.S. users. This fee, which covers a portion of our telecommunication charges for Uninet access, is your total cost for the BIONET Resource. Aside from the manuals, there are no other charges for this service.

Foreign investigators, including Canadians, on the BIONET system will be responsible for supporting their own network communications.

Sincerely,

Mary Lou Warner
Administrator
BIONET

APPLICATION CHECKLIST

__Provided user information (page 1)

__Completed INTENDED USE OF BIONET and current grant support statement (page 3)

__Marked DRR Scientific Classifications (page 4)

__BIONET User agreement read and signed by Principal Investigator and responsible grant administrative officer (page 6)

__Filled out documentation order form (if desired)

__Copy made for your records.

Mail the completed application to:

BIONET Application
IntelliGenetics, Inc.
1975 El Camino Real West
Mountain View, CA 94040

Incomplete applications cannot be processed and will be returned. Please send all inquiries about this application to the above address. Include your name, phone number and application date in all correspondence.

Applications are processed once a month. The cut-off date is the 20th. Applications received on or after that date will be processed the following month.

Application Form for the BIONET[tm] Resource

See reverse for description and eligibility of user classifications

Date of Application:

Principal Investigator (full name and title):

Affiliation: Department, School and Institution:

Mailing Address (Include a Street Address for parcels shipped UPS):

Area code and phone number:

Applying for Class I, II, III or IV Use?: _____ (See Reverse for more information)

Would you like more information on the BIONET Satellite program?: _____ (See Reverse for more information)

Type of terminal or terminal emulator to be used:
(example: Tektronix 4023 or IBM-PC with VT100 emulator)

Type of communications software to be used:
(example: KERMIT or Smarterm) (note: KERMIT, a public domain communications software, is available from our lending library. Please indicate if you would like to borrow a disk for copying.)

Additional users: List up to 5 BIONET users under your direction. (Each Principal Investigator is allocated a fixed amount of space on the computer and only one user in a group can be logged in at one time.) Highlight the primary contact person for your group if not yourself. *NAC acceptance rules require individual qualifying PI's to apply separately. Additional PI's listed on this application will not be given access.*

Name	Title	Phone	Position

Criteria for Eligibility

The four classes of user status are described below. Most users will be Class I users or IV users. Please call or write if you would like to be considered for Class II or III status.

CLASS I: Researchers from academic and non-profit institutions who can demonstrate that they are supported by governmental, philanthropic, or unrestricted institution funds and that their research can be assisted by the resource facilities. Exceptions will be considered on a case-by-case basis.

These users will have access to the programs in the Core, Database, and Contributed Libraries, and to the electronic mail and bulletin board facilities.

An annual fee of $400 is charged for this access. *PI's in foreign countries, including Canada, will not pay the subscription fee but must pay their own telecommunication costs.*

CLASS II: Scientists who wish to participate in developing the BIONET Resource by providing new programs to the community. Acceptance as a Class II user is determined in part by the relevance of their programs. These programs should help achieve the goals described in the cover letter.

Class II user must meet the same eligibility requirements as the Class I users. However, they will also receive support from the BIONET staff in developing and making their contributed software accessible to the Bionet community. The Class II user will not be required to pay the subscription fee.

Please include a description, in detail, of what you intend to contribute, what support you will need from the resource and how the work will benefit the BIONET community. Also include a list of current publications in the area of intended use (for the past two years only).

CLASS III: People responsible for Department, School or Campus-wide computer facilities who wish to provide information about or access to BIONET to the community they serve. These users must provide evidence of their position and responsibilities for providing computer facilities for a local community of scientists with access to BIONET.

CLASS IV: Scientists who wish to take advantage of only the electronic communications facilities - electronic mail, bulletin boards, and file transfer programs - will be given restricted access for an annual fee of $100. These users must meet the eligibility requirements of the Class I user.

BIONET SATELLITE PROGRAM

In addition to the above classes, BIONET, in cooperation with IntelliGenetics, is now able to offer an on-site BIONET package. Utilizing existing Digital Equipment VAX or 2060 computers, or SUN Microsystems, all of the programs, bulletin boards and electronic mail functions would be accessible at your location. Your local scientific community would benefit by having a direct and more powerful access to the resource. Accessing this service requires the purchase of a software license from IntelliGenetics. A special purchase program has been arranged to make it easy for academic institutions to join the BIONET Satellite program. If you are interested, please contact us directly or mark the appropriate response on the reverse.

Intended use of BIONET. Include a Research Title of 80 characters or less, and a Research Abstract with a minimum of 3 lines and a maximum of 350 characters. Class II and III applicants, in addition, should include additional information described in the Criteria for Eligibility on page 2. *You may attach a separate sheet if you prefer.*

Current grant support in area of intended use. Include each federal grant by Principal Investigator, title, funding institution, grant number and duration of support and a brief (three to ten line) abstract of the research. If funding is from institutional or other unrestricted funds, provide information on sources of funding sufficient for the NAC to determine if conditions for access have been met. *If this funding is scheduled to end within 12 months, state whether a renewal of the same grant/funding is pending.*

Appendix to Instructions - DRR Scientific Classification

AXIS I Code Nos.	Resource Material/Research Area (Maximum 4 Codes)	AXIS II Code Nos.	Research Areas (Maximum 4 Codes)
1	Animals:	30	Aging
	a. Vertebrates, Mammal	32	Anesthesiology
	b. Vertebrates, Non-Mammal	34	Anthropology/Ethnography
	c. Invertebrates	36	Behavioral Sci/Psychology/Social Sci
2	Biological/Chemical Compounds	38	Bioethics
3	Biomaterials	40	Communication Science
4	Cells & Subcellular Material	42	Computer Science
5	Human Subjects	44	Congenital Defects or Malformations
6	Membrane/Tissue/Isolated Organ	46	Degenerative Disorders
7	Microoganisms:	48	Device Prothesis Intra/Extracorporea
	a. Bacteria	50	Drug Studies:
	b. Virus		a. Toxic c. Orphan Drugs
	c. Parasites		b. Other
	d. Other	52	Engineering/Bioengineering
8	Plants/Fungus	54	Environmental Sciences:
9	Technology/Technique Development	54	a. Toxic
10	Other (SPECIFY)		b. Other
12	Clinical Trials:	56	Epidemiology
	a. Multicenter b. Single Center	58	Genetics, Including Metabolic Errors
		60	Growth and Development
ANATOMICAL SYSTEM/RESEARCH AREAS		62	Health Care Applications
		64	Immunology and Allergy
13	Cardiovascular System	66	Infectious Diseases
14	Connective Tissue	68	Information Science
15	Endocrine System	70	Instrument Development
16	Gastrointestinal System:	72	Mental Disorders/Psychiatry
	a. Esophagus	74	Metabolism and Transport:
	b. Gallbladder		a. Carbohydrate
	c. Intestine		b. Electrolyte & Water Balance
	d. Liver		c. Enzymes
	e. Pancreas		d. Gases
17	Hematological System		e. Hormone
18	Integumentary System		f. Lipid
19	Lymphatic and Recticulo-Endothelial System		g. Nucleic Acid
			h. Protein & Amino Acid
20	Muscular System	76	Neoplasms/Oncology:
21	Nervous System		a. Benign
22	Oral/Dental		b. Malignant
23	Reproductive System	78	Nutrition
24	Respiratory System	80	Radiology/Radiation Nuclear Medicine:
25	Sensory System:		a. Ionizing (Xray, Nuclear Reactor)
	a. Ear		b. Non-ionizing (Microwave, Radar)
	b. Eye	82	Rehabilitation
	c. Taste/Smell/Touch	84	Statistics/Mathematics
26	Skeletal System	86	Surgery
27	Urinary System	88	Transplantation
28	Other (SPECIFY)	90	Trauma
		92	Other (SPECIFY)

BIONET^{tm} User Agreement

- The BIONET resource will not be used for any commercial purpose which is not specifically identified to and approved by BIONET's National Advisory Committee (NAC). Any pertinent change in sponsorship, continuity of grant support, or use made of BIONET will be reported promptly to the BIONET Resource Manager.

- The NAC will approve all access and will make the final judgment on applications that are questionable in nature, scope, or funding of research.

- Standard DEC-2060 facilities for file protection will be available to protect the integrity of your data and programs.

- Ownership of data and software developed on or contributed to the Resource will be subject to the guidelines of the Principal Investigator's institution and granting agency, to which all questions on legal issues should be directed. The BIONET Resource will retain a non-exclusive, royalty-free right to use, by approved BIONET investigators, of the data and executable versions of the software on BIONET.

- All investigators granted BIONET access must provide brief annual summaries of research results. The summaries must be included in our annual report of Resource activites to the NIH. Investigators will have sufficient advance notice to prepare the summaries.

- All publications that involve use of the Resource must acknowledge the Resource by name and NIH grant number (e.g.: *Computer resources used to carry out our studies were provided by the BIONET^{tm} National Computer Resource for Molecular Biology, whose funding is provided by the Biomedical Research Technology Program, Division of Research Resources, National Institutes of Health, Grant #1 U41 RR-01685.*) Investigators must send three (3) copies of these publications to the Resource Manager.

- Access to BIONET will be granted to a Principal Investigator and designated members of his or her research group. Each group will be allocated a fixed amount of disk storage space distributed by the PI and designated associates. Class II users will be granted larger amounts of disk space.

- We request that each PI limit access of his or her group to one login to BIONET at a time. Use of the Resource will be carefully monitored by the staff and the NAC.

- The BIONET Resource provides only a computer facility and associated services. It does not provide research equipment. The Resource has a small fund for fostering collaborations and will use this fund, when no other means are available, to support an effort that will advance the goals of the Resource.

I assume full responsibility for all users listed on this application form and will monitor their compliance to the conditions and restrictions for access to the BIONET Resource. I will inform the BIONET Consultant, (electronic mail address BIONET), by electronic mail, immediately about any changes in this group of users, i.e., departure of an existing user or addition of new staff qualified to use the Resource. I will inform new users of the above mentioned conditions and restrictions.

As Principal Investigator of this grant to use the BIONET Resource, I agree, by signing this application, to adhere to all conditions and restrictions for use of the BIONET Resource, as described above and such further regulations as may be issued from time to time by the NIH or the NAC.

Signature of Principal Investigator: _____

Date: _____

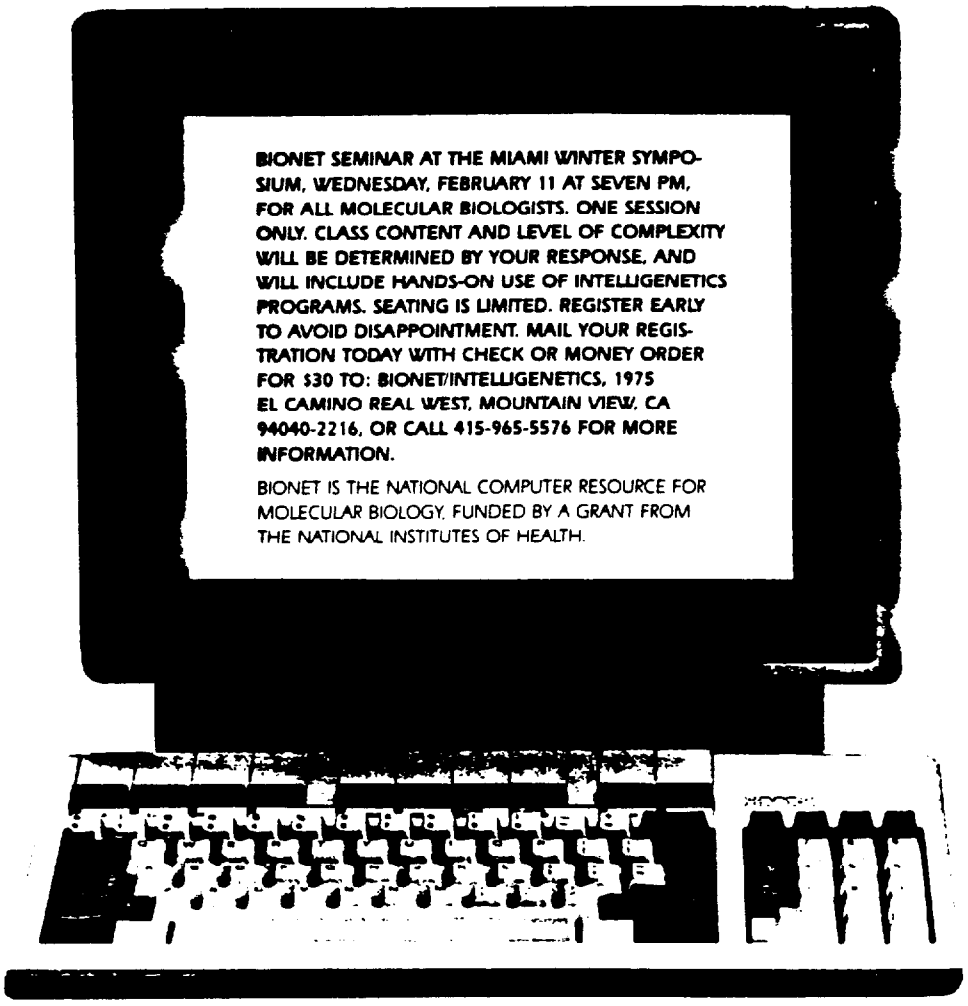
I have also furnished a copy of this application to the responsible grant administrative officer of my institution, whose name and signature are given below:

Name of official: _____

Signature: _____

IX. ADVERTISMENT FOR BIONET TRAINING SESSION

If you are a MOLECULAR BIOLOGIST you may be eligible to join BIONET™...



Clip or copy, and mail with check or money order for $30 to BIONET/IntelliGenetics, 1975 El Camino Real West, Mountain View, CA 94040-2216

Investigator _____

Institution _____

Address _____

City _____ **State** _____ **Zip** _____

☐ **New user** ☐ **Advanced user**

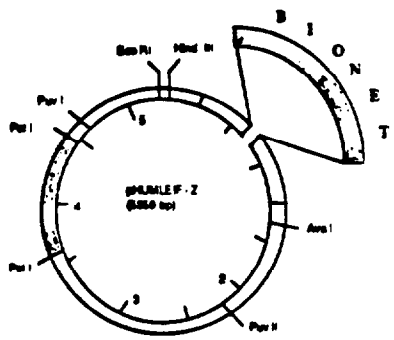
☐ **I am interested in hosting a session at my institution**

Please check topics of interest to you

- ☐ Managing large DNA sequencing projects
- ☐ Restriction mapping tools
- ☐ Simulation and design of recombinant DNA experiments
- ☐ DNA or protein sequence database organization & search methods
- ☐ DNA or protein sequence analysis
- ☐ Sequence comparison methods

X. RENEWAL NEWSLETTER

DO YOU KNOW WHAT YOU'RE MISSING?



SECOND YEAR OF BIONET IS GREAT SUCCESS!!!

ITS TIME TO RENEW YOUR SUBSCRIPTION NOW!

BIONET has entered its third year stronger than ever. In keeping with its projected schedule, the BIONET staff has:

- added more databases
- brought in contributed software
- expanded the Bulletin Boards
- upgraded existing programs
- doubled the number of telecommunication ports
- established a training program

and more . . . WITHOUT INCREASING THE SUBSCRIPTION FEE.

We are excited about the growth and changes in BIONET and hope you are too. For more information on any of the above enhancements to the Resource, please give us a call. . . or better yet, log-in and check it out yourself!

ANNOUNCING . . .

INTELLIGENETICS BECOMES JOINT VENTURE

The Amoco Corporation of Chicago has purchased a controlling interest in IntelliGenetics from IntelliCorp, Inc., of Mountain View, making IntelliGenetics a venture jointly owned by the two companies. IntelliGenetics will continue to market its current line of molecular biology programs and will maintain its traditional emphasis on customer support.

This relationship with Amoco will provide IntelliGenetics with greater resources for the development of new software. There are plans to add several new programs to the software that runs on the SUN workstation, the VAX minicomputer, the microVAX II, and the timesharing system.

The most sophisticated new software will be Strategene, a genetic engineering workstation based on artificial intelligence technology.

Molecular biologists at the Amoco Research Center have been working with knowledge engineers at IntelliCorp for the past two years to apply IntelliCorp's KEE™, an integrated package of AI tools, to problem solving in molecular biology. The success of this work led to the formation of the joint venture with IntelliCorp.

Equipped with a mouse and windows, Strategene lets scientists rapidly simulate complex cloning experiments in a graphical environment. The accessibility of DNA information and the ease and accuracy of simulations make it possible for scientists to experiment with a much larger number of vectors than they would ordinarily use.

Strategene uses AI techniques to organize knowledge about DNA molecules. This knowledge encompasses both descriptive information and rules for reasoning about cloning experiments. The system contains a reference library of vectors and allows researchers to enter and retrieve information from individual and laboratory libraries of constructions.

Strategene is designed to operate in conjunction with IntelliGenetics' package of analytic software. The system currently runs on a Xerox 1186.

GET YOUR UPDATED INTRODUCTION TO BIONET FREE WITH YOUR SUBSCRIPTION RENEWAL. ACT NOW!

INTELLIGENETICS ANNOUNCES. . .

PC/GENE

A Personal Computer Genetic Engineering Environment

PC GENE is a comprehensive package of molecular biology software for microcomputers. It contains almost fifty different programs for analyzing peptides and nucleic acids.

You can use PC/GENE as the perfect companion to BIONET, or you can run it independently. Data can be transferred efficiently through a modem connection. This allows you to perform large database searches and sequence comparisons on BIONET. At the same time you can take advantage of PC convenience and graphics capabilities to run a host of different analyses in your own laboratory.

Thus, BIONET subscribers can still get the speed and memory of a large computer when they need it.

PC/GENE microcomputer software comes with the same high level of support you have come to expect from IntelliGenetics. Our scientific account representatives will offer the same degree of personal service for this new software package.

PC/GENE allows biologists with little computer experience to use the programs productively in a matter of minutes. The system presents a series of choices of analyses that are expressed in terms that biologists use. It is necessary only to click the mouse or press a single key to choose a sequence to analyze, to define parameters, or to display the results in a variety of ways.

ALL UNINET DIAL-UP PHONE NUMBERS CHANGING IN SEPTEMBER

Uninet is being absorbed into GTE Telenet to form US Sprint Telenet. This means that the phone numbers to access the BIONET central resource will change. Additionally, we regret that there will be a light change in the procedure used once you dial up. This change will occur in September. Each PI will receive a special mailing in August with all the details. Information will also be available on BIONET via the sign-on banner.

As a positive benefit of this combined network, US Sprint Telenet will have access numbers in over 50 new local dialing areas. A database of access numbers is available on Telenet to all users.

We are working to arrange a two week overlap when both the old and the new access methods will work. The Uninet dial-ups will be in service for 6 weeks following the change, but BIONET will not be accessible through them. Please consider this if you will be out of touch with BIONET for a month or more. Starting in September, access via the old UNINET dial-ups will produce an error message.

We will try to make the transition as smooth as possible. In the event of any problems reaching BIONET electronically, you can telephone the consultant at 415 324-GENE for assistance.

Some of the analyses that PC/GENE performs on peptides are:

- Computing best oligonucleotide probe
- Predicting antigenic determinants
- Searching for peptide subsequences
- Comparing sequences using the Needleman Wunsch algorithm
- Aligning two sequences
- Determining secondary structure using the Chou and Fasman or the Garnier method
- Predicting membrane associated alpha helices
- Plotting local concentrations of amino acids
- Calculating statistics of usage of di and tripeptides
- Plotting a protein's hydropathic index

Some of the analyses that PC/GENE performs on nucleic acids are:

- Displaying tRNA in a clover leaf configuration
- Translating sequences
- Translating introns and exons using EMBL database annotations
- Searching for subsequences in nucleic acids
- Searching for coding regions using both Fickett's and Shepherd's methods
- Finding restriction sites, altering restriction enzymes lists, digesting sequences
- Creating a restriction site with a single mutation
- Comparing sequences with the Pustell dot matrix method
- Searching for hairpin loops
- Analyzing nucleic acid sequence statistics: codon usage, local base concentrations, enriched sequences

BIONET™ National Computer Resource for Molecular Biology is funded through a cooperative agreement with IntelliGenetics, Inc., by the Biomedical Research Technology Program, Division of Research Resources, National Institutes of Heath. Grant Number RR01685

IntelliGenetics, Inc. is located at 1975 El Camino Real West, Mountain View, CA 94040. Phone 415 965-5575

ELECTRONIC BULLETIN BOARDS COME ALIVE!

by Nancy Bigham

One of the three major goals of the BIONET Resource is to promote collaboration and rapid sharing of information among a national community of scientists. The BIONET bulletin boards (bboard) meet this goal by providing a facility whereby BIONET users can exchange data, laboratory techinques and ideas with others in similiar fields.

During the past 6 months, the BIONET Resource has been concentrating on updating and improving the bulletin boards. There are now 29 bulletin boards - containing articles ranging from reviews of pc-communications software to an article about Fast Fourier Transforms and related algorithms for sequence analysis (MOLECULAR-EVOLUTION, message #12).

Prominent members of the BIONET community have been selected to be bulletin board leaders. They will provide the bulletin boards with the most recent and pertinent information. Under this new leadership, a dynamic bulletin board community is being developed by encouraging a lively interchange of information, maintaining a vital resource of community news, and archiving outdated bulletins.

With the participation of the community and the work of the bboard leaders, the bboards contain more exciting and pertinent information than ever before. Please take time to view the bboards in your field of interest and to contribute information to any of the bboards. For more information about reading the bboards and contributing messages, see your INTRODUCTION TO BIONET manual.

CURRENT BULLETIN BOARD LEADERS

Gene-Expression	William Sofer
Genomic-Organization	Steve Harris
Libraries	Larry Kedes
Molecular-Evolution	Dan Davison
PC-Software	Doug Brutlag
Plant-Molecular-Biology	Ronald Sederoff
Politics	Michelle Cimbala
RNA-Folding	Michael Zuker

VECTORBANK UPDATED By Ellen Hartzler

VectorBank is IntelliGenetics' collection of maps of common vectors designed for use in the CLONER program.

It is easier to use because we have added two new files. VECTORBANK.LST is a list of all vectors available in VectorBank, and VECTORBANK.TXT is a description of some important features of VectorBank. As in previous releases of VectorBank, we have provided multiple maps for each vector. The following list of maps for PBR322 illustrates the file naming convention.

File name	Map Description:
PBR322_6CUT.MAP	All six-cutter sites.
PBR322_COH.MAP	All cohesive cutter sites.
PBR322_COM.MAP	All prototype sites.
PBR322_FLUSH.MAP	All flush cutter sites.
PBR322_UNQ.MAP	All unique cutter sites.

Please note that we have replaced the hyphen in the all names with an underscore, i.e., PBR322-6CUT.MAP is now PBR322_6CUT.MAP.

To obtain a vector map for use in CLONER, follow these procedures:

1. Find the vector you want by looking at VECTORBANK.LST. This is a listing of all the vectors in VectorBank.
2. Enter the CLONER program and LOAD your vector from VectorBank.

If you are unfamiliar with the CLONER program, work through the tutorial on CLONER in your BIONET training manual.

NEW TRANSLATION TABLES IN SEQ

by Terry Friedemann

A new addition to SEQ provides two different ways to alter the codon tables used in SEQ for translating a DNA sequence.

- You can directly edit the codon table that contains the standard genetic code with your particular codon changes and save those changes.
- Or, if you use the translation tables supplied by the program, you can translate sequences using a different genetic code without having to do any editing.

To see examples for yeast mitochondria codon changes with both the new editable codon table option and one of the new translation options, log on and send your request to BIONET using Electronic Mail.

EVOLUTIONARY ORIGIN OF HEPATITIS B VIRUS AND RETROVIRUSES

by MaryJo Lawler

Dr. William Robinson, a Professor at Stanford University and one of the first researchers to join BIONET, and Roger Miller, a postdoctoral fellow, have used BIONET in their research on the molecular structure of Hepatitis B virus.

Initially, Dr. Robinson and Dr. Miller chose to study the secondary structure of the origin of replication of known hepadna viruses. In examining stable palindromes near the origin of replication, they discovered by computer manipulation that the regions flanking these palindromes were highly conserved.

In an effort to substantiate these findings, they performed several global searches through the Genbank and EMBL databases and found that not only were these regions conserved across hepadna viruses, but they were also present in type C retroviruses. As a result of further analyses in the lab, they gathered additional evidence which led them to suggest that HBV and retroviruses have a common evolutionary origin.

The authors have stated that without the use of computer analysis software and access to a complete and up-to-date database, the question of the genetic relatedness of the hepadna virus and retrovirus families would not have been raised.

The investigators retrieved the hepadna virus sequences from the Genbank database using the QUEST program. They performed the initial DNA homology and palindrome analyses with the SEARCH function of the SEQ program, and employed other SEQ commands to examine base-composition and to translate sequences.

Drs. Robinson and Miller discovered that the regions were conserved in 27 viral DNA sequences by searching over the Genbank and EMBL databases using the IFIND program. Additional searching using IFIND and the SEARCH functionality of PEP demonstrated a high degree of homology between the HBV core protein and the retroviral P30 gag nucleocapsid protein. The investigators also used the PEP program for open reading frame analysis, hydropathicity plots and secondary structure prediction.

Dr. Miller made extensive use of the electronic mail and bulletin board facilities on BIONET to trade unpublished hepadna virus sequences with several other labs on the system.

COMPUTERS HELP ANALYZE SHOPE FIBROMA GENE by MaryJo Lawler

Grant McFadden and Chris Upton, working at the University of Alberta, have used the BIONET Resource extensively for their work on the molecular organization of the Shope fibroma gene. They have submitted for publication a paper entitled "DNA Sequence Homology between the Terminal Inverted Repeats of Shope Fibroma Virus and an Endogenous Cellular Plasmid Species."

In their paper, Dr. McFadden and Dr. Upton discuss three research findings and suggest how they correlate: the presence of an extrachromosomal autonomous DNA species, its hybridization to Shope fibroma virus (SFV), and the exchange of genetic information between host cells and cytoplasmically replicating poxviruses. The investigators used BIONET exclusively for their computer analysis.

They used the GEL program extensively for sequence entry and assembly. Once supplied with the sequence, Dr. Upton used the SEQ program to study the inverted repeat regions of the SFV DNA and to analyze the cytoplasmic DNA molecules through restriction enzyme, base composition, open reading frame, and translation analyses.

The investigators' homology comparisons were done using the SEARCH function of the SEQ program. Additional homology searches using IFIND over the Genbank and EMBL databases showed no additional homologous sequences to the inverted repeat region of SFV, but revealed similarity between

continued on page 6

PEP'S DIGEST OPTION

DIGEST is a major new addition to the protein analysis functions in PEP which is designed to help you study proteins by rapidly simulating the action of a peptide digestion with proteases or by chemicals.

The program provides a list of commonly and not so commonly used proteases and cleavage chemicals. You can add to this list, or you can create an entirely different list. When you add a new protease, DIGEST allows you to place the cleavage site before or after the recognition site. DIGEST also accommodates proteases that cleave at more than one site.

Once you are satisfied with the list of proteases, you can run DIGEST, choosing one or more proteases from the list. The resulting digestion simulation shows the location, length, and molecular weights of the fragments.

There are several additional options open to you. You can ask to see a map of the cleavage sites or see amino acid composition data for the fragments. You can treat any of the fragments as if they were independent peptides and then analyze them with any of the PEP functions. You can also ask DIGEST to simulate a peptide fingerprint by asking the program to draw a plot of the molecular weight of the fragments versus their isoelectric points.

As in all IntelliGenetics programs, you can ask for on-line help. Before you begin the option, we recommend that you read the introduction after the DIGEST: prompt.

SIMPLE SEARCHES

by Doug Brutlag and Alan Engelberg

The computer operating systems from which you use IntelliGenetics programs provide several tools on the DEC 2060 for rapidly searching unformatted databases and text files such as sequence data files.

On the DEC 2060 the fastest and simplest tool is the FIND program which is good for looking for one or a few patterns in a single file. The FIND program has the SCOPE concept from QUEST in that it allows you to look for a pattern in a line, a paragraph, or a page. It permits a limited amount of ambiguity but only allows you to examine a single file at a time. Using FIND is analogous to looking in a phone book for a person's name.

The simplest and most common application of FIND is to type FIND WITHIN <line> <pattern> IN <filename>, *(see example below)* leaving out all the other qualifiers.

continued on page 7

Predicting Experimental Results

When you perform a restriction digestion of a newly cloned sequence and electrophorese the resulting fragments, CLONER can save a great deal of time by quickly and accurately predicting the possible digestion patterns. In the following brief example we show how you can determine the orientation of your clone. If your vector contains more than one potential insertion site, you can use the same procedure to determine into which site you've cloned the insert.

CLONER: load pbr322+com.map *We will insert our fragment into pBR322.*
Reading file PBR322_COM.MAP ...
1. PBR322-COM (4363 N) C ; DEFINITION PLASMID PBR322 (E.COLI CLONING VECTOR)
CLONER: new
NEW allows us to enter the restriction information about the insert. If we had sequence information, we could create a restriction map in SEQ and then load that map into CLONER.
Name for new map: PragZ
Length of new map: 1480
Topology: linear
Enter as many new lines of comments as desired; End with an extra <CR>
; Contains GeneZ
; <CR>
Please enter each site name followed by its location.
Finish with a blank entry.
Site name and cut position(s): pstI 1 1480
Site name and cut position(s): hamhI 245 750
Site name and cut position(s): ecori 1200
Site name and cut position(s): <CR>
PragZ is map number 2.
Loading editor help text...
MapEdit: *region*

continued on page 6

@FIND WITHIN line actin, myosin IN nih.lst<CR>
FIND shows each line where a hit occurs.
ACAACT1 ;AMOEBA (A. CASTELLANII) ACTIN GENE-1.
BOVACT1 ;BOVINE ACTIN MRNA, 5' END.
BOVACT2 ;BOVINE ACTIN MRNA, 3' END.
BOVPRL ;Bovine prolactin (prl) mRNA.
Since FIND simply searching for a sequence of characters, it will report hits when that sequence appears in a larger sequence, i.e., it finds a hit on actin in the word prolactin.
BOVPRLP1 ;BOVINE PROLACTIN, 5' FLANK AND EXON 1.
BOVPRLP2 ;BOVINE PROLACTIN, 5' FLANK AND PARTIAL EXON 2.
CELACT1 ;CAENORHABDITIS ELEGANS (NEMATODE) ACTIN I GENE 5' END.
CELACTII ;CAENORHABDITIS ELEGANS (NEMATODE) ACTIN II GENE 5' END.
CELACTIII ;CAENORHABDITIS ELEGANS (NEMATODE) ACTIN III GENE 5' END.
CELACTIV1 ;CAENORHABDITIS ELEGANS (NEMATODE) ACTIN IV GENE 5' END(SEG 1).
CELACTIV2 ;CAENORHABDITIS ELEGANS (NEMATODE) ACTIN IV GENE 5' END(SEG 2).
CELMYH ;C.ELEGANS MAJOR MYOSIN HEAVY CHAIN (UNC-54 I) GENE, 3' END.
Here FIND reports a hit on the second pattern, myosin
CELMYUNC ;C.ELEGANS MAJOR MYOSIN HEAVY CHAIN ISOZYME UNC-54 I GENE

If you want to see the name of the file where the sequence is located, you simply leave out "within line" in the FIND command line. The default scope is paragraph and that makes it possible to see the file name.
@find actin, myosin in nih.lst
*ACA.NIH
> ACAACT1 ;AMOEBA (A. CASTELLANII) ACTIN GENE-1.
* ACARRS8S ;A.CASTELLANII (AMOEBA) 5.8S RIBOSOMAL RNA.
* ACARRSS ;A.CASTELLANII (AMOEBA) 5S RIBOSOMAL RNA.

*BOV.NIH
> BOVACT1 ;BOVINE ACTIN MRNA, 5' END.
* BOVACT2 ;BOVINE ACTIN MRNA, 3' END.
The pointer ">" indicates the line with the matching string of letters.

PREDICTING continued from page 5

Name for new region: GeneZ
Region boundaries: 140 1320
Fill character: (<CR>=-) <CR>
Polarity (<, |, or >): (<CR>=|) >
Region GeneZ is now on level 1
MapEdit: quit
CLONER: list
1. PBR322-COM (4363 N) C ; DEFINITION PLASMID PBR322 (E.COLI CLONING VECTOR)
2. FragZ (1480 N) L ; Contains GeneZ
CLONER: insert 2 1 pstI *We simulate the insertion of FragZ into pBR.*
Name for new map: (<CR>=PBR322-COM-FragZ)
Retain comments from PBR322-COM? (Y, N, D, ?, or ^) (<CR>=Y) no
Retain comments from FragZ? (Y, N, D, ?, or ^) (<CR>=Y) <CR>
Enter as many new lines of comments as desired; End with an extra <CR>
;This map is FragZ inserted into pBR at the pstI site.
;<CR>
PBR322-COM-FragZ is map number 3.
CLONER: edit 2
MapEdit: flip *Flipping the map of the insert will allow us to simulate a fragment inserted with the reverse orientation.*
Area to invert: all
MapEdit: quit
CLONER: insert 2 1 pstI *We repeat the same insertion but in this map the orientation of the insert is reversed.*
Name for new map: (<CR>=PBR322-COM-FragZ) pbrZ-flipped
Retain comments from PBR322-COM? (Y, N, D, ?, or ^) (<CR>=Y) no
Retain comments from FragZ? (Y, N, D, ?, or ^) (<CR>=Y) no
Enter as many new lines of comments as desired; End with an extra <CR>
; FragZ inserted in pBR in the opposite direction to map PBR322-COM-FRAGZ.
; <CR>
pbrZ-flipped is map number 4.

To determine the orientation of the insert, we simulate a digestion with the enzyme chosen to analyze the clones and see which digestion matches the experimental results. We could have run simulated digestions with a number of enzymes to see which would give us the most distinct results.

CLONER: digest 3 bamhi

Enzyme	Site	Length	Enzyme	Site
BamHI	(376)	3482	BamHI	(3858)
BamHI	(4363)	1856	BamHI	(376)
BamHI	(3858)	505	BamHI	(4363)

CLONER: digest 4 bamhi

Enzyme	Site	Length	Enzyme	Site
BamHI	(376)	3968	BamHI	(4344)
BamHI	(4849)	1370	BamHI	(376)
BamHI	(4344)		BamHI	(4849)

After we have electrophoresed the restriction digest fragments we need only compare the gel pattern to the two sets of patterns above to determine the orientation of the insert.

SHOPE continued from page 4

the extracellular DNA and a family of cellular protease inhibitors.

In addition to having access to analytical programs, Dr. Upton is pleased with the opportunity to use electronic mail to communicate with other scientists. Like many other BIONET scientists, Dr. Upton had little computer experience prior to BIONET. He has since become very active in the bulletin board communities. Dr. Upton has traded codon usage tables with several other BIONET scientists and has become one of the community's MacIntosh authorities. He is currently working with an investigator in New York, whom he met through interactions on BIONET, and they are setting up what they call a "personal network" for their collective analysis needs. He foresees using BIONET even more extensively than in the past, especially because the McFadden lab has sequenced 15 to 20KB since he began work in the group.

DID YOU KNOW

When you are editing a meld, you can save your edits in three different ways.

- "SAVE gels" is the program default. Editing changes you have made are retained for the current session only and have not been written in the .pro file. If you lose your job either because the computer crashes or because there are telecommunications problems, then you will lose those edits.

- "SAVE files" saves them permanently. These edits cannot be lost.

- "Set autosave on" will automatically save your editing changes in the .pro file if you type "Set autosave on" after the "Medit:" prompt after the "Medit:" prompt.

SEARCHES cont. from page 5

(type FIND<cr> to see a description of these qualifers). For example, NIH.LST is a Genbank file that contains a one line entry for each Genbank sequence. On this line appears the sequence name and the first line of comments. If you were searching for a word or two that you expected to appear in the definition line, you would search the file NIH.LST as shown on page 4.

FIND can also determine whether VectorBank contains a particular vector. You can search the file vectorbank.lst, a list of all the vectorbank maps. To see if pUC is present, type FIND WITHIN line puc IN vectorbank.lst.

Many people are using QUEST to search for simple unambiguous keys in sequences or in comments. A much faster and simpler program called XSEARCH (see example below) will allow you to search databases for keys with no ambiguous letters.

To run the program, type XSEARCH after the "@" prompt. XSEARCH is not as convenient as QUEST in that you cannot COLLECT hits nor can you control the output. However, it searches databases 10 to 50 times faster than QUEST and is useful for an initial screen if you don't need ambiguous bases. Once XSEARCH has reported the names of the files that contain exact hits you can use QUEST to search just these files and then COLLECT or print out the

SEARCHES cont. page 8

```
@XSEARCH
 SUBSTRING search routine (compiled 11-Jul-80) ? for help
Files to search: @nih-primate.fls<CR>
Files to search: (continued) : <CR>
Target 1) myosin<CR>
Target 2) actin<CR>          You can search for more than one pattern.
Target 3) <CR>
Equivalences:  1// <CR>
current expression: 1 V 2        V = or, i.e. 1 or 2 is noted as a hit.
Expression:
Create .PL files? NO//<CR>
Output goes to: * TTY: // <CR>      This sends the output to your terminal.
Type DEL or RUBOUT to abort any particular file search.
Searching <SEQUENCES>APE.NIH.8510
Searching <SEQUENCES>GCR.NIH.8510
Searching <SEQUENCES>HUM.NIH.8510
Searching <SEQUENCES>HUMA.NIH.8510
Searching <SEQUENCES>HUMA1.NIH.8510
Searching <SEQUENCES>HUMAC.NIH.8510
When XSEARCH finds a match it displays the file in which the match is
located and the line in which the hit occurs.
(<SEQUENCES>HUMAC.NIH.8510 1.1) {actin}
; DEFINITION  HUMAN BETA-ACTIN RELATED PSEUDOGENE H-BETA-AC-PSI-1 5'END.
(<SEQUENCES>HUMAC.NIH.8510 1.4) {actin}
; KEYWORDS    ACTIN; PROCESSED GENE.
(<SEQUENCES>HUMAC.NIH.8510 1.12) {actin}
;  TITLE     STRUCTURE OF TWO HUMAN BETA-ACTIN-RELATED PROCESSED GENES ONE
OF
(<SEQUENCES>HUMAC.NIH.8510 1.19) {actin}
;              SITE        420  1540       HOMOLOGOUS TO ACTIN READING FRAME
(<SEQUENCES>HUMAC.NIH.8510 2.1) {actin}
; DEFINITION  HUMAN BETA-ACTIN RELATED PSEUDOGENE H-BETA-AC-PSI-1 3'END.
(<SEQUENCES>HUMAC.NIH.8510 2.4) {actin}
; KEYWORDS    ACTIN; PROCESSED GENE.
    .
    .
Searching <SEQUENCES>HUMMY
Searching <SEQUENCES>HUMTB.NIH.8510
Searching <SEQUENCES>HUMTR.NIH.8510
(<SEQUENCES>HUMTR.NIH.8510 1.1) {myosin}   Here is a hit with another target pattern.
; DEFINITION  HUMAN NON-MUSCLE (FIBROBLAST) TROPOMYOSIN GENE.
          Lines recognized = 190
  String    Matches  Unrecognized Matches
1) "myosin"         3          0
2) "actin"          205        0
Letter case ignored ("Ab" = "aB").
Files with no matches: <SEQUENCES>APE.NIH.8510, <SEQUENCES>GCR.NIH.8510, ... <S
EQUENCES>MNKR.NIH.8510.
68 files searched, 63 without matches.
^L
          ...DONE...  continue to start over
```

FINDING NEAR-RECOGNITION SITES WITH QUEST by Jaya Carl

QUEST's flexibility makes it possible to search for a great variety of patterns in sequences. For example, QUEST can be used to design a key to locate sequences of bases that are one base away from being a restriction enzyme site and that, if changed, would not alter the translation of the sequence. The purpose of this search is to locate a place to introduce a new recognition site to easily identify positive clones.

In the keys described below we have developed patterns that search for a set of near-EcoRI sites, but the same procedure can be used to find any near-restriction site that does not alter the translation.

Since we don't want to alter the translation we must determine the frame in which we are making the change. Thus the first part of the key is:

A[TU]G & (...){1,}

This key represents the MET start codon immediately followed by one or more triplets.

In order not to alter the translation of the sequence when we alter the single base that introduces the recognition site, we must take advantage of the degeneracy of the genetic code. The recognition site for EcoRI is GAATTC. We can make a change in the third base of a codon. The reading frame determines which base is the degenerate one. The first key is:

ATG & (...){1,} & GAGTTC.

In this key, the reading frame is such that the first G is the first base of a codon. The third base was changed from A to G because both of these codons code for Glu.

The next key is:

ATG & (...){1,} & GAATTT.

In this key the reading frame is the same as the one above except that we are making a change in the second codon, changing the codon from TTC to TTT, since both code for Phe.

However, there is no reason to assume this particular reading frame with regard to the recognition site. Instead of there being an even set of triplets between the start codon and the recognition site, there could be one or two additional bases. For one additional base the key is:

ATG & (...){1,} & (.) & GAATCC.

In this case the reading frame is shifted by one, so we want to look for gaATCc instead of gaATTc, since both ATC and ATT code for Ile.

The final key is:

ATG & (...){1,} & (..) & GAACTC.

This key assumes the frame shift is 2.

For an example of the way to use this key, simply log on and send your request to BIONET, using the Electronic Mail.

IntelliGenetics also maintains a database of key patterns that you can use in QUEST to help identify various structural and consensus regions in nucleic acid and protein sequences. The files are located in the <IG> directory.

We have collected the following files. If you have written a useful key, we would be delighted to include it in the KeyBank library.

AA.KEY	Identifies codons for antigenic sites.
AACOMP.KEY	Identifies codons of complementary strand for antigenic sites
AMINO.KEY	Equates one-letter amino acid code with three-letter code.
GENE.KEY	Identifies open reading frames.
KEY1.KEY	Shows keys from Quest Help Topic KEY1-EXAMPLE.
KEY2.KEY	Shows keys from Quest Help Topic KEY2-EXAMPLE.
KEY3.KEY	Shows keys from Quest Help Topic KEY3-EXAMPLE.
KEY4.KEY	Shows keys from Quest Help Topic KEY4-EXAMPLE.
NAD.KEY	Identifies dinucleotide-binding region for peptides.
OPIOID.KEY	Identifies potential DNA encoding endogenous opioid activity.
PROMOTER.KEY	Shows suggested consensus sequences for procaryotic promoters.
REST.KEY	Identifies prototype restriction enzyme recognition sequences.
SIGNAL.KEY	Identifies consensus for leader peptide cleavage site.
ZDNA.KEY	Shows potential Z-DNA purine-pyrimidine pattern.

SEARCHES cont. from page 7

CONTEXT around the hits. XSEARCH first prompts you for "Files to Search" and you may respond with filenames containing wildcards or indirect filenames (@NIH-PRIMATES.FLS). Then it prompts you for "Targets" which are just character strings to search for. If you specify more than one target XSEARCH then prompts you for a Boolean relationship between the targets and the default is to search for target 1 OR target 2 OR target 3 OR ...

XSEARCH next asks you for equivalences and the default (obtained by hitting carriage return) is to equate upper and lower case letters. If you wish to have XSEARCH search through sequence information rather than comments then you should type the letter A (with NO carriage return!) at the "Equivalences:" prompt, and when it asks you for an "equivalence file," type SEQUENCE.XSE. This file not only equates upper and lower case, it equates T's and U's and causes XSEARCH to ignore carriage returns, line feeds, tabs and other punctuation in sequences. It is equivalent to SEQUENCE SCOPE in QUEST. Otherwise XSEARCH works exclusively in LINE SCOPE. Try XSEARCH. Type a ? (with NO carriage return) at each prompt to find out much more about XSEARCH's capabilities and limitations.

XI. BULLETIN BOARD LEADER AD

HOW TO SAVE $400 AND HELP BRING YOUR FIELD INTO THE COMPUTER AGE

The BIONET-NEWS bulletin board has messages posted which describe the variety of uses for the bulletin board system and file transfer facilities. These uses range from having a continuous on-line scientific meeting in your research area to sending manuscripts to colleagues in distant labs. The list could undoubtedly be extended by creative people. (See also HELP MEETINGS.)

To encourage expanded use of the communications facilities, particularly the bulletin board network, we are offering a

FREE ONE YEAR BIONET SUBSCRIPTION

to users who are willing to organize and lead a bulletin board. Bulletin board leaders should be actively engaged in research in the selected area of interest.

Leading a bulletin board would involve contributing items of interest to the board, encouraging other people in the research field to participate (leaders should have plenty of contacts!), monitoring incoming messages, archiving dated material, and finally submitting a brief year-end report on the bulletin board activity to BIONET. Renewal of the position would be subject to a yearly review by BIONET. We estimate that the work involved would occupy only a few hours each month, but some responsibilities could be delegated to other lab members.

Prospective leaders should submit a proposal via electronic mail to BIONET. The proposal should include a description of the suggested bulletin board along with an estimate of the number and potential activity of participants. The activity estimate could be gathered by e-mail contacts prior to submitting the proposal. The final selection of bulletin board topics and leaders will be made in conjunction with BIONET and its National Advisory Committee. Please contact your BIONET consultant at 415-324-4363 if you have any questions.

A list of current bulletin board topics and names of leaders can be obtained by typing HELP BB-LIST after the prompt. Some of the current boards need leaders. However, new topics are especially encouraged.

As more molecular biologists and biochemists become computer-literate, participation in the bulletin board system should accelerate. As activity increases, the leadership positions will grow in influence. This is your opportunity to get involved with a new communications medium at its inception!

Somebody will eventually lead your research field into the computer age. Why not make it you?