BIONET: An NIH Computer Resource for Molecular Biology

Douglas L. Brutlag and David Kristofferson
BIONET
700 East El Camino Real
Mountain View, CA 94040, USA

Introduction

BIONET is a computer network for molecular biologists and is run as a non-profit resource for the scientific community by IntelliGenetics, Inc. BIONET is funded by the National Institutes of Health (NIH), Division of Research Resources, and provides scientists access to numerous biological sequence databases and to software tools required to analyze these data (Smith et al., 1986; Kristofferson, 1987). Another major goal is to provide a means for electronic communications between scientists to encourage collaborations. BIONET also serves as a central facility for development and distribution of novel software in the area of molecular biology and develops methods for distributed computing.

In addition to the two major nucleic acid sequence databases of GenBank and the European Molecular Biology Laboratory (EMBL) DNA database, BIONET provides access to the Protein Identification Resource (PIR) protein sequence database, the Cold Spring Harbor restriction enzyme database, the Brookhaven Protein Structure Database, the VectorBank of restriction maps of vectors, phages and viruses, a compendium of known oncogenes, and Lindsley and Grell's Variations of *Drosophila melanogaster*. Thus, on this single resource we have information at several levels of biological function immediately available to researchers.

The service component of BIONET has been successful as judged by the more than 500 laboratories using the facility and the nearly 2,000 scientists using the software. BIONET's primary attraction is the ready

availability of well supported software and databases. Molecular biologists routinely use BIONET to compare their most recent DNA or protein sequences against the currently available sequences. While 1,000 complete searches are processed each month on BIONET, nearly 3,000 searches of subsets of the database are also performed. The other analytical software is accessed nearly 10,000 times per month by various BIONET users. Its 28 communications ports are very often full at peak hours, and the BIONET community uses considerably more than the 50 percent of the CPU cycles of a DEC 2065 allotted to it by the BIONET Grant.

Communication

With connections to three national communications networks and linkages to many other international networks, BIONET is readily accessible to scientists. BIONET is linked directly with TELENET and CompuServe which provide 24-hour access to the BIONET computer via a local telephone call from most cities in the United States. These networks also make BIONET available to scientists in Europe and Japan via international carriers and networks such as EURONET, Datax-P, TransPac and Venus. BIONET is also directly connected to ARPANET, which provides mail and file transfer capabilities between any host computer on the ARPANET, and to a large number of INTERNET hosts on BITNET, USENET, CSNET, etc. This high level of connectivity has facilitated communications and collaboration among scientists worldwide.

Collaboration

BIONET has established a library of user-contributed software that can be accessed directly on the BIONET computer and much of which can be downloaded to microcomputers. These programs are maintained on the BIONET computer by the original developers, who are often assisted by BIONET staff in making their programs compatible with the BIONET computer and file formats. These programs (and their developers) include:

- ALIGN and PC-ALIGN, to quickly align DNA sequences (K. Thompson and Dan Davison)

- BIOFOLD and PC-FOLD, to minimize free energies in RNA structures (Michael Zuker)

- DM, a suite of sequence analysis programs (David Mount)

- FASTP and FASTN, for rapid database searches (Bill Pearson and David Lipman)

- GEL, which sizes DNA fragments (John Thompson)

- IDEAS, a suite of programs for DNA and protein sequence comparisons (Minoru Kanehisa)

- MOLECULE, which displays RNA structures graphically (John Thompson)

- PROT3, which gives the best alignment of three protein sequences (Joel Sussman)

- SEQAIDII, a multifunctional program for DNA sequence analysis (Douglas Rhoads and Don Roufa)

- XMULTAN, to align many sequences simultaneously (Bill Bains)

- XPROF, for calculating hydropathic profiles of proteins (George Rose)

A number of additional programs, primarily involved in multiple sequence alignment and the calculation of sequence phylogenies, have been contributed by others and are currently being adapted to the BIONET environment.

Electronic Bulletin Boards

The communications provided by BIONET have also aided the community by encouraging collaboration and sharing of data and pro-

grams. A number of bulletin boards on scientific topics are available to the BIONET community, among them, Gene-Expression, Genomic-Organization, Methods-and-Reagents, Molecular-Evolution, Onco-genes, Plant-Molecular-Biology, Protein-Analysis and Yeast-Genetics.

Recently, BIONET has set up an important exchange of bulletins with the SEQNET bulletin board managed by Martin Bishop at MRC Laboratories in Cambridge, England. SEQNET is a molecular biology bulletin board distributed on the JANET academic network and linking many universities in the United Kingdom. BIONET users can post messages directly by mailing bulletins to SEQNET from the BIONET computer. Similarly, molecular biologists in the United Kingdom can post messages on the BIONET bulletin boards by including BIONET in their address list.

Finally, a new set of bulletin boards permits BIONET users to communicate with the developers of a series of commercially available software packages. These bulletin boards provide a site for discussion of the strengths, weaknesses and improvements of these packages and direct electronic communications with the developers and support staff. These new bulletin boards include the software packages Mac-Gene, PC/Gene, MicroGenie, IBI/Pustell, the Arizona Genetic Software Programs and the University of Wisconsin programs.

Links with Other Resources

Since its beginning, BIONET has maintained electronic communications with several other resources, including the NIH GenBank, the EMBL data library, the Molecular Biology Computer Research Resource (MBCRR) and the Protein Identification Resource (PIR). Initially this took the form of electronic mailboxes for representatives of these resources on BIONET and reciprocally at these other sites. Recently we have established a GenBank bulletin board so that the BIONET community can provide comments and feedback directly to this valuable resource.

Because a great number of DNA sequences were actually being determined using the IntelliGenetics GEL program on the BIONET computer, we felt that BIONET could provide a further service to the community by developing software that would allow scientists to annotate their sequences according to the standard GenBank format

and mail those sequences and their annotation to GenBank electronically. The GenPub program, as it is called, is a forms-oriented display editor that allows individuals to fill in a template based on the GenBank submission form (and which can be readily changed if the GenBank form changes) giving all the requisite data about a sequence. The program automatically inserts the sequence in the appropriate place in the form by copying the sequence from a designated file on the BIONET computer. When the form is completed, a single keystroke forwards the information via ARPANET to the GenBank computer at Los Alamos for inclusion in the next issue of the database. At that point, the entry is verified by the GenBank staff and if they have questions about the data, they can query the author by electronic mail at BIONET.

By having direct electronic submission, GenPub eliminates many errors in transcription. GenPub takes sequence files directly from the software used to determine the sequence and submits it to the database. If local PC software is used to determine a sequence, the data can be sent to BIONET using an error-checking protocol such as Kermit or Modem and forwarded to GenBank via GenPub, eliminating all transcription errors. More importantly, GenPub recruits the scientist who has determined a sequence to annotate it, thus eliminating the problems of reading and interpreting the publication, and further simplifying the job of sequence collection.

Future BIONET Development

It was clear from the outset that a primary role of BIONET as a resource of CPU cycles and databases for biological sequence analyses was untenable. No computer resource, however powerful, could ever fulfill the computing needs of the molecular genetics community. A second problem with a central resource is the limited communications bandwidth and hence response time from the central computer.

At its inception, BIONET was designed as a network to foster communications and distributed computing. BIONET provides distributed computing in several ways. First, by encouraging the distribution of PC-based programs with important functionality for the molecular biologist, it moves much of the routine computing to a local environment. Second, BIONET has developed special batch versions of the

core IntelliGenetics software that allow either later job submissions on the same machine, or submission to another machine. Fortunately the programs that are most CPU intensive (sequence database searching) are also very amenable to non-interactive batch file submissions. Even batch file submissions at off-peak hour• have resulted in much more effective use of the BIONET computer. Batch files submitted to other resources will soon double the capacity of the BIONET resource at a minimal cost. These same procedures will also allow BIONET to distribute jobs to other facilities which have excess capacity and, perhaps, even to super computer resources.

The final mechanism fostering distributed computing involves the BIONET Satellite program. This program provides remote usage of the BIONET software tools at other institutions, and allows researchers to use a system which they are most familiar with and over which they have more direct control than a central facility. Many users prefer accessing a local computer resource than a remote one, if only because of the higher communication bandwidth with that resource. To maintain the critical connectivity of BIONET Satellites with other BIONET users, we have developed special mail and bulletin board communication programs that allow scientists at BIONET satellites to send mail to BIONET and post bulletins from their local computer. These communications can be facilitated if the BIONET Satellite computer is an Internet host; however, the BIONET communications software will work on a host with only a single dedicated modem.

In addition to these developments in BIONET computing, new additions to the BIONET staff in the last year have enabled us to improve our user-support service to the community and undertake a variety of projects to enhance the resource. Our best measure of success will be the extent to which we make the powerful features of the BIONET resource accessible to the large numbers of researchers who have a minimal background in computing. A good start has been made in this direction with the provision of well-organized and clear on-line help. New on-line examples of program usage and improvements in the overall program user interface are also underway. Reaction from the community to these new features has been extremely positive, but changes planned for the near future will dwarf these accomplishments.

Too often, computing involves hours, even days, of specialized training before one can begin to work productively. Ease of use is

often attained only by sacrificing powerful features. Our goal is to develop a powerful, yet "friendly" computer system, one that can be used easily by every scientist, not just by a small, highly-trained group of specialists. In the next year BIONET will make significant progress towards realizing that goal by providing new means of interfacing with its software.

References

Kristofferson, D. (1987). The BIONET electronic network. *Nature,* 325, 555-6.

Smith, D.M., Brutlag, D.L, Friedland, P. and Kedes, L.M. (1986). BIONET: National computer resource for molecular biology. *Nucleic Acids Research,* 14, 17-20.