The molecular biology computer research resource

T.F.Smith, K.Gruskin¹, S.Tolman² and D.Faulkner²

Harvard School of Public Health, Boston, MA, ¹Wayne State University, Detroit, MI, and
²Dana-Farber Cancer Institute, Boston, MA, USA

Received 16 August 1985

ABSTRACT

Described is a new National Institutes of Health supported Molecular Biology Computer Research Resource located at the Dana-Farber Cancer Institute in association with Harvard University, Boston, Massachusetts.

INTRODUCTION

The National Institutes of Health, Division of Research Resources has awarded a 5 year grant to the Dana-Farber Cancer Institute, Harvard School of Public Health and Harvard Medical School to establish a molecular biology computer resource coupled with a genetic sequence pattern recognition program. This Molecular Biology Computer Research Resource, the MBCRR, has two major collaborative objectives. The first is research in the development of statistical and computer methods that can reveal syntactic and semantic patterns in DNA, RNA, and protein sequences--most of which are still obscure or unknown. The other is to support a research resource to use such methods--developed in the course of this project as well as others--to discover patterns that trace evolution and that are involved in the regulation of gene expression. The identification of such sequences should provoke the formulation and testing of major hypotheses in molecular evolution, gene regulation and developmental genetics.

Temple F. Smith of the Harvard School of Public Health is the director of the MBCRR. Co-principal investigators with Smith are Allan M. Maxam of Harvard Medical School and Marcello Pagano of the Harvard School of Public Health. Other MBCRR staff are Susan Tolman, User Coordinator*, Donald V. Faulkner and Ronald Newman, Application Programmers.

Allan Maxam was responsible for initially bringing data bases and software of interest to the local molecular biology community at the Harvard Medical

* Karen Gruskin who was instrumental in the start of the MBCRR has left as user coordinator to continue her education, but remains a consultant.

Nucleic Acids Research

School four years ago. This acted as a pilot program for the MBCRR proposal to the NIH. The MBCRR has been organized to exploit the advantages of coupling active basic genetic research directly with computer software development. This is to be accomplished by encouraging collaborative interactions between active researchers who have new or unsolved analysis problems and computer scientists and software developers.

COMPUTER RESOURCE

The MBCRR provides numerous DNA, RNA, and protein sequence analysis tools to investigators in the greater Boston area on-line and to the research community at large by downloading or on magnetic media. The MBCRR maintains a number of data bases on-line and is in the process of obtaining access to others. Currently included are nucleic acid sequence data bases GenBank and EMBL (European Molecular Biology Laboratory), the NBRF (National Biomedical Research Foundation) protein sequence data base, the Kabat "Proteins of Immunological Interest" data base, the Cold Spring Harbor Restriction Enzyme data base and the IntelliGenetics cloning vector data base. In addition the MBCRR is working to make available some of the biological data bases not yet accessible and to provide a system to enable cross referencing with the current sequence data. A front end (in FORTRAN77 and "C") has been developed allowing access to the data base which may be added to most software. The MBCRR has developed a utility to reformat and subdivide (important for small computer systems) the original GenBank releases into user defined files. With the rapidly expanding data base size and large number of alternate analysis programs, the MBCRR has written utilities to browse through on-line documentation, help files and program example run files and data base directories. All such developed tools are freely available.

TRAINING

The MBCRR provides training/research opportunities at three levels. First there is introductory level training in the use of available software (including BioNet (1) software) at the Dana-Farber Cancer Institute. This introduction is generally used only by local area users, but is available to any researcher visiting the area. Second, more extensive training with collaborative research opportunities will begin in April 1986 with visiting scientist positions available for one to three months. These positions are designed to give hands-on computer software experience to molecular biologists who have current research projects requiring computer analysis. The MBCRR

will also work to develop new analysis approaches as needed. It is presumed that upon return to their own institutions these scientists will be able to act as local experts to introduce others to the use of available software. Finally, beginning in April 1986, the resource will offer training to molecular biologists in statistics and computer methods, and to biomathematicians and computer scientists in contemporary molecular genetics by providing interdisciplinary postdoctoral fellowships.

In addition, the MBCRR and the Theoretical Biology and Biophysics Group at Los Alamos National Laboratories (responsible for GENBANK) will co-sponsor an international symposium on computer analysis in molecular genetics in 1986. Collaboration with other NIE resources such as BioNet and the NBRF protein data base is planned for this symposium.

RESEARCH

The basic research currently under way is centered on the problems of genetic sequence pattern recognition. A number of algorithmic approaches to identify and define such important regulatory patterns as promoters, enhancers, and splice sites are being investigated. What is needed is a method that will identify the most characteristic pattern for any functional region. One method under investigation is related to the estimation of probability densities of Parzen (2). In order to identify those patterns most distinctive of a particular region a Chi Square statistic was calculated for each potential pattern occurrence distribution histogram. With each position one associates the largest Chi Square value for the occurrence histogram that had its maximum deviation at that position. Thus relatively large associated values indicate regions for which a particular pattern is not only shared by a large number of the sequences but is distinctive of that region. Identified using this approach on a set of *E. coli* promoters are the previously recognized -10 and -35 regions, and two new regions; one just 5-prime of the aligned start sites and another around -43.

A consensus sequence approach under study collaboratively with Walter Fitch of Wisconsin and Morris Goodman of Wayne State University involves iterative dynamic programming. Here, a standard dynamic programming algorithm (3,4) is used to generate alignments of each sequence in the set with the statistically weighted consensus from the previous iteration.

The MBCRR has developed a number of graphic display programs for working with protein secondary structure predictions. These include a high information density color display of the charge, hydrophobicity, and

structural feature predictions, which can be coupled with a real-time folding alpha carbon back bone three-dimensional display. Ideally, one hopes to combine such secondary structure information with amino acid chemical group equivalences and minimum mutation induced sequence alignments. While there is no single computer program that combines all such information into a single sequence comparative analysis, combinations of a wide variety of software can allow the effective layering of ones biological insight and experience with mathematical analyses.

DISTRIBUTION

With the recent advances in computer hardware and the reduced costs of personal computers, many researchers have purchased their own computer. To aid in the transfer of routine computer analysis to the many micro computers, the MBCRR has obtained permission to distribute a number of software packages designed especially for use on personal computers:

- 1) Mount (Univ. of Arizona) Sequence Utilities (5).
- 2) Fristensky (Cornell Univ.) DNA Sequence Utilities.
- 3) Lipman & Pearson (National Institutes of Health) Rapid Biosequence Similarity Analysis System (6).
- 4) Zucker (Canadian National Research Council) RNA Secondary Structure Prediction Code (7).
- 5) RZMAP (Ralph. N. Mich. Univ., & Fitch, Univ. of Wisc.) Single/Double Digest Restriction Map Code (8).
- 6) PROSTR (Ralph, N. Mich Univ., & Smith, MBCRR) A Chou-Fasman utilizing protein secondary structure prediction code incorporating hysteresis.
- 7) ALIGN (Fitch, Univ. of Wisconsin, & Smith, MBCRR) A generalized dynamic programming pairwise sequence alignment code (4).
- 8) DASHER (Faulkner & Smith, MBCRR) A high speed GenBank similarity search code.
- 9) RUNMAT (Smith, MBCRR) A shared similarity run matrix code (replacing the older dot matrix codes) incorporating statistical significance.

The MBCRR has agreed to distribute the source, the documentation and, MS-DOS executables of specified codes. The MBCRR welcomes the opportunity to review and consider for distribution software developed by any molecular biologist. This should relieve many researchers who have developed useful computer tools of the problems of distribution inquiries and documentation

reproduction. The MBCRR will, if agreed upon, collaborate with such software developers in updating and/or adding needed front end modifications prior to distribution.

LONG TERM GOALS

The MBCRR plans to expand its on-line services, its accessible data bases, and its library of programs and utilities. Anticipated is an increase in both the on-line user community and the greater national community's exploitation of the available training and software distribution facilities. These expansions are linked to the planned MBCRR accessibility over major computer networks. The latter will also encourage expansion of existing collaborations including those through BioNet. The current wealth of genetic sequence information is phenomenal and expanding rapidly. The sheer size of this and other related data will allow for the investigation of new emergent properties, some of which will be discovered from theoretical explorations using new computer analyses. The MBCRR plans to be an active partner in such research.

ACKNOWLEDGEMENTS

The authors would like to thank the Biostatistics Department of Harvard School of Public Health and Dana-Farber Cancer Institute for support prior to NIH funding. This resource is supported by NIH Grant, RR02275 and (TFS, DVF) in part by GM35677.

REFERENCES

1. Lewin, R. (1984). Science, 223, 1379-1380.
2. Parzen, E. (1962). Ann. Math. Statist., 33, 1065-1076.
3. Smith, T.F., Waterman, M.S. (1981). Adv. in Appl. Math., 2, 482-489.
4. Fitch, W.M., Smith, T.F. (1983). PNAS, 80, 1382-1386.
5. Mount, D., Conrad, B. (1984). Nucl. Acids Res., 12, 819-824.
6. Wilbur, W.J., Lipman, D.J. (1983). PNAS, 80, 726-730.
7. Zucker, M., Stiegler, P. (1981). Nucl. Acids Res., 9, 133-148.
8. Fitch, W.M., Smith T.F., Ralph W.W. (1983). Gene, 22, 19-29.