****
* MacroMolecules, Genes, and Computers Symposium *
* and Matrix Workshop 1986 *
* REPORT *
****

An unusually wide range of biological problems were discussed at two recent meetings held in Waterville Valley, N.H. under the sponsorship of the MBCRR of the Dana-Farber Cancer Institute and Harvard University. They ranged from the integration and access of all known biological and related computer databases, to the problems of genetic sequence syntactic pattern identification. Along the way the problems involved in the mapping and sequencing of the entire E. coli and human genomes, the standardization of molecular databases, the mathematics of sequence comparisons, the prediction of molecular structure and the reconstruction of molecular evolution were all addressed.

The first of these meetings, the Matrix meeting, was a two day summer workshop planning session organized by James Willett (NIH), Harold Morowitz (Yale) and Karen Gruskin (Wayne State) through the MBCRR. It was funded by the Biological Models and Materials Resources Section, ARP, DRR NIH. This was an outgrowth from an earlier study by the National Research Council's committee on models in biology, which had introduced the concept of a global "matrix of biological information". The task of this session was to begin planning for a 1987 summer workshop on the feasibility of using the current tools of computer science and the wealth of existing databases to bring this matrix concept to a useful accessible embodiment. One is not envisioning a new single restructured database of all biological information, rather it is assumed that with proper indexing, thesaurus development, and sophisticated multi-database intelligent interfaces, coordinate database access and searching will be possible. • This is inline with the recent National Research Council recommendations to the National Library of Medicine in a report from the May 5-6, 1986 meeting.

The idea for the 1987 workshop is to bring together a reasonable number of bright young scientists under expert leadership with two major objectives: the first being to introduce the attendants to the methodologies of Artificial Intelligence and computer science required to interface, cross reference and search for inherent structural principles within the existing and developing biological information databases; the second being to investigate a number of research subject areas where solutions appear to require information and new insights from across the expanse of the matrix of biological information for their resolution. The identification analogous and homologous systems (and models) across the wide diversity of living organisms and understanding their evolution are obvious problems expected to be aided by such an approach. In addition, part of what is to be investigated, is whether truly new insights and perspectives are possible given real multi database

accessibility and hands-on use of the most powerful search, analysis and graphic computer tools.

The second meeting was coordinated by Temple Smith (Harvard\DFCI) and Karen Gruskin (Wayne State) both of the MBCRR and was funded in part by the Biomedical Resource Technology Program of DRR, NIH. The MacroMolecules, Genes and Computers symposium not only demonstrated that molecular biology encompasses most of the traditional divisions of classical biology, but that for those researchers generating and interpreting the new genetic sequence information a very broad range of expertise is required. This includes everything from computer science and evolutionary biology to population genetics and structural biochemistry.

The vast majority of the presented papers were overviews of the various areas with an emphasis on the remaining analysis problems. Dr. Walter Gilbert (Harvard) in the keynote address help set the tone for the meeting: one by opening discussions on the feasibility and utility of mapping the entire human genome, and secondly by noting that the exon-intron organization of the eukaryotic genes suggests an alternate organizational structure for the sequence databases. These databases might be organized as sublibraries of functional or structural domain units, each containing the modern day representatives of the original functional genetic domains used to assemble the great diversity of known proteins and regulatory units. Discussions on the importance of mapping the human genome continued throughout the meeting. It was considered of the utmost importance to generate a reliable physical map and cosmid library of the human genome which would be correlated with all known genetic, disease and sequence information.

A number of pilot projects are already under way employing everything from pedigree analysis to restriction site polymorphisms and oligo- nucleotide annealing distributions. Discussed were the X-chromosomal mapping at Strasbourg, the EMBL linking library work, and the RFLP (restriction fragment polymorphisms) mapping at Yale and the Whitehead Institute with Collaborative Research Inc. It was pointed out that the levels of polymorphisms, incomplete pedigrees, the distribution of highly repetitive sequences and pseudo genes frequencies are all complicating these mapping efforts. It is worthy to note that the Wisconsin group under Fred Blattner announced that nearly forty percent of E. coli has been sequenced and that they are nearing a complete physical mapping of its genome. In relationship to this and other large sequencing projects, GenBank and EMBL announced collaborative work is under way to update the structure of the nucleic acid sequence databases. This will include a standardization of the annotation. It was however recognized that with the anticipated increase in sequencing efforts and the changing view as to what are the important features within the sequences, a major restructuring of the molecular sequence databases may be required to allow for both their effort use and growth.

Some concerns were raised that as major human sequencing projects expand over the next few years, that monies and effort may be drawn from related cross species comparative sequencing studies. It was felt that only through such studies will the relative importance of sequence polymorphisms and conserved homologies be understood. The latter, are perhaps, essential for the proper identification of the very short and generally degenerate genetic regulatory signal sequences.

With much of the meeting's emphasis on analytical methods it was a little surprising that no major analysis breakthroughs were announced, rather it seemed that there was a clear division of the problems into the very hard, whose solutions will require fundamental mathematical and/or biochemical insights, and those which seem only to require careful and expanded application of already known and employed methodologies. In the former category are of course the problems of predicting function and structure from primary sequence information including the identification of new regulatory patterns and protein functions. While powerful statistical, heuristic and graphical tools have been usefully employed, a general computer science or mathematical theory of genetic syntax or functional genetic pattern recognition seems a fair ways off. Both the mathematicians Samual Karlin (Stanford) and Michael Waterman (USC) noted that numerous generalizations of the currently employed methods will no doubt prove helpful in moving toward such a theory.

Also in this area of difficult problems are many of the problems associated with evolutionary reconstructions. Many of these are now known to be NP-complete mathematical problems, and thus often require the use of good heuristics. Here it seemed that the various incongruities in both the data and the methodologies may be our best sources of new insights as pointed out in the discussions on the discovery of gene conversions by Ben Koop (Wayne State) and later by Joseph Felsenstein (U of Washington) in discussing tree building algorithms. There are cases where the shear weight of the data may be sufficient, as discussed in the studies of hemoglobin evolution (Ben Koop) and t-RNA evolution (Walter Fitch, USC).

In the latter category of less difficult problems, the implementation of current methods including new computer hardware is making the molecular biology bench top workstation a reality. These workstations have software to organize, characterize and compare new sequence information, including tools for identifying homologies and well known functional sites shared with sequences in the sequence databases. They also provide access to powerful graphic displays available until recently only on mainframe systems at major centers. A number of groups are exploring the use of parallel processing for high speed implementation of the rigorous dynamic programming sequence alignment algorithms. A Princeton group (represented by Douglas Welch) has already tested a new chip designed especially for this problem. While problems remain, particularly in the speed and efficiency of many

algorithms and in the organization and representations of the rapidly expanding sequence information, no fundamental understandings appear to stand in the way of providing very powerful analysis and search packages on bench top workstations. This may turn out to be overly optimistic as the scale of the mapping and sequencing of the E. coli and human genome progress over the next decade or so. Yet as demonstrated in the three afternoon workshops, much of the routine analysis and molecular data management software is currently available in a wide variety of commercial and academic packages. Available are a wide range of user interfaces, databases and graphic tools with most available on a range of hardware.

The meeting was well attended by a little over two hundred researchers in computer science, mathematics and molecular and evolutionary biology. Over thirty national, international and regional molecular database and computer resource centers participated including GenBank, BioNet, the NBRF-PIR, EMBL, and the MBCRR. (See following lists) In addition there were representatives from all the major US funding agencies including the Department of Energy, the National Science Foundation, the National Institutes of Health and the National Library of Medicine.

Under the sponsorship of the Alfred P. Sloan Foundation and contributions from the attending commercial companies (see list below) travel and other support was made available to a large number of young and foreign scientists. The exposure of so many younger members of this scientific community, particularly those with mathematical and computer science backgrounds, to the active researchers in molecular genetics was perhaps the meeting's most important success. In order to further foster such interactions at this rather large meeting, a very successful series of small dinner-discussion parties became one of the highlights, lasting late into one of the evenings.

The organizers of these meetings would like to thank all of the participants for helping to make them a success. In particular, I would like to thank: George Bell, Fred Blattner, Roy Britten, Georges Cohen, Charles Coulter, Gerald Fasman, Don Faulkner, Jonathan Fineberg, Walter Fitch, Walter Gilbert, William Gilbert, Walter Goad, Karen Gruskin, Bob Langridge, Nancy Ludtke, Allan Maxam, Harold Morowitz, William Ralph, Dan Shaffer, Gary Stormo, Susan Tolman, Sherman Weissman, Susan Wheeler, Jim Willett, Tracie Uliss, for their help and support.

This report submitted by Temple F. Smith, Director MBCRR, 8/29/86
Karen Gruskin, Meeting Coordinator

NATIONAL AND REGIONAL RESEARCH COMPUTER RESOURCES PARTICIPATING IN THE MACROMOLECULES, GENES AND COMPUTERS SYMPOSIUM AND WORKSHOP

BIONET

COMPUTER GRAPHICS LABORATORY
UCSF

COLD SPRING HARBOR

CORNELL SEQUENCE ANALYSIS PACKAGE

CSIRO, North Ryde, Australia

EMBL DATA LIBRARY

GENBANK

GIRS, A Einstein College of Med.

GENETICS PC - SOFTWARE CENTER
Univ of Arizona

GENEUS

HUMAN GENE MAPPING LIBRARY

HUMAN GENE PROBES/CHROMOSOME
LIBRARIES DATA BANK

HYBRIDOMA DATA BANK

MICROBIAL STRAIN DATA NETWORK

MODEL AND DATABASE COORDINATION
LABORATORY, Dept of Agriculture

MOLECULAR BIOLOGY COMPUTER
RESEARCH RESOURCE, DFCI

MBIR, Baylor College of Medicine

NATIONAL CENTRE FOR BIOINFORMATICS

NATIONAL LIBRARY OF MEDICINE

NEWAT PACKAGE UCSD

ILL. NATURAL HISTORY SURVEY DIV.
PHYLOGENETIC ANALYSIS

PROPHET

PROTEIN DATA BANK, Brookhaven

PROTEIN IDENTIFICATION RESOURCE

ROCKEFELLER UNIV.

SEQUENCE ANALYSIS SYSTEM OF
THE CLINICAL RESEARCH INSTITUTE

SEQUENCE ANALYSIS SYSTEM OF
THE INSTITUT PASTEUR

UCSF - BIOMATHEMATICS COMPUTATION
LABORATORY

UNIVERSITY OF WISCONSIN GENETICS
COMPUTER GROUP

WHITAKER COLLEGE COMPUTER FACILITY

WHITEHEAD INSTITUTE

COMMERCIAL COMPANIES WHO CONTRIBUTED FINANCIAL AND/OR EQUIPMENT SUPPORT FOR THE MACROMOLECULES, GENES AND COMPUTERS SYMPOSIUM AND WORKSHOP

Applied Biosystems

Battelle,
Pacific Northwest Division

Cambridge Scientific Abstracts

Computer Signal Processing Inc.

Digital Equipment Corporation

DNAstar, Inc.

Evans & Sutherland

GraphOn

ImClone Systems Incorporated

Intelligenetics Incorporated

International Biotechnologies, Inc.

International Business Machines

IRL Press

Mary Ann Liebert, Inc. Publishers

NEC Information Systems, Inc.

Silicon Graphics

Stockton Press

Sun Microsystems

Tektronix Inc.

TEXTCO

TRIPOS Associates, Inc.