IDENTIFYING FEATURES IN BIOLOGICAL SEQUENCES: SIXTH WORKSHOP REPORT

AUTHOR(S):
Christian Burks, T-10
Eugene Myers, Univ. of Arizona
William R. Pearson, Univ. of Virginia

RECEIVED
MAR 13 1996
OSTI

Theoretical Biology and Biophysics Group (T-10)
MS K710, T-10, Los Alamos, NM 87545
Identifying Features in Biological Sequences:  
Sixth Workshop Report  
(Aspen Center for Physics, 29 May - 16 June, 1995)

Christian Burks¹  
Eugene Myers²  
William R. Pearson³

Los Alamos National Laboratory  

¹Theoretical Biology and Biophysics Group  
T-10, MS K710  
Los Alamos National Laboratory  
Los Alamos, NM 87532  
U.S.A.

e-mail: cb@t10.lanl.gov  
fax: 505-665-3493  
tel: 505-667-6683

²Dept of Computer Science  
University of Arizona

³Dept of Biochemistry  
University of Virginia
abstract

This report covers the sixth of an annual series of workshops held at the Aspen Center for Physics concentrating particularly on the identification of features in DNA sequence, and more broadly on related topics in computational molecular biology. Over the last six years the workshop has been cited for inspiring or otherwise contributing to fifty-five papers, has provided training and inspiration to Ph.D students and post-doctoral fellows early in their careers, and has provided senior scientists the unique opportunity to seriously engage in interdisciplinary collaborations over the three weeks of interactive residency.

overview

The Aspen Center for Physics (ACP), in Aspen, Colorado, sponsored a three-week workshop in May-June 1995, with twenty-five scientists participating, fourteen for their first time. The workshop, entitled Identifying Features in Biological Sequences, was the sixth in an annual series hosted by ACP on the general topic of assembling, comparing, and identifying features in biological sequences. The previous workshops occurred in years 1990 through 1994.1-4

The workshop series originally focused primarily on discussion of current needs and future strategies for identifying and predicting the presence of complex functional units on sequenced, but otherwise uncharacterized, genomic DNA. We addressed the need for computationally-based, automatic tools for synthesizing available data about individual consensus sequences and local compositional patterns into the composite objects (e.g., genes) that are -- as composite entities -- the true object of interest when scanning DNA sequences. The general background and justification for a workshop on this topic was discussed in the first workshop report.1 Of particular interest over the past few years has been the maturation of previously described as well as the emergence of several new approaches to predicting the presence and location of genes.2-8

The workshop was structured to promote sustained informal contact and exchange of expertise between molecular biologists, computer scientists, and mathematicians. No participant stayed for less than one week, and most attended for two or three weeks. Computers, software, and databases were available for use as ‘electronic blackboards’ and as the basis for collaborative exploration of ideas being discussed and developed at the workshop.

There have been no recent (or even not-so-recent) meetings devoted to precisely the topic that provided the theme of our workshop. Though there have been a number of workshops over the years devoted to DNA sequence analysis, none have focused on the recognition and characterization of composite objects (such a genes) exclusively; for this reason, this workshop series has provided a unique approach to addressing a very important challenge in making use of the data coming out of large-scale sequencing projects.

There are very few meetings that:

• last for several weeks,
include scientists from disparate disciplines,
and provide a structure promoting informal interaction and primary emphasis on pairwise (or small group) intensive exchange of ideas and insights.

As with previous years, this workshop provided an unusual and very facilitating environment, and time for scientists to address the necessary and considerable learning curves associated with unfamiliar disciplines. Several long-term formal and informal collaborative interdisciplinary projects resulted from this workshop that would not otherwise have developed, partly because of the mixture of scientists from different disciplines and partly because of the sustained interaction that the workshop afforded.

The educational value of the workshop is worth noting. Exchange of information, and particularly description of problems to those most able to provide the appropriate tools were a central theme of this workshop. Furthermore, since several of the participants were post-doctoral fellows or graduate students, the workshop provided interdisciplinary training that will presumably be very useful to them, even -- in some cases -- influencing them in deciding to pursue one of the interdisciplinary research paths.

*Dates.* The workshop extended from 29 May to 16 June, 1995.

*Location.* The ACP (Aspen, Colorado) hosted the workshop, and provided facilities and administrative support staff, including an Administrative Vice-President and four full-time secretaries on site for responding to housing, word-processing, and secretarial requests. ACP provided offices for use by participants, with two scientists per office. Several lecture/conference rooms with projection capabilities and seating capacity for a group of our size were available (and used), including an outside patio conference room for informal seminars and discussions. Condominium style residences were provided (through the ACP) that allow for routine meal preparation, informal gatherings, and family/guests. A more detailed description of ACP and its support for workshops such as ours was given in the first workshop report.

*Organizing Committee.* C. Burks, E. Myers, and W. Pearson constituted the Organizing Committee, with C. Burks serving as Chairperson, for the workshop’s technical agenda. T. Appelquist, as President of the Aspen Center for Physics, was the formal contact for administrative aspects of the workshop.

*Funding.* Workshop support was provided by a standing NSF grant (#PHY-9104428) to ACP for its summer program, and from a grant of $15K specific to this year’s workshop from DOE’s Office of Health and Environmental Research. The DOE/OHER funds were used to partially offset the participants’ costs in attending the workshop, to set up the workstation network, and to provide general administrative support for the workshop, consistent with the general approach used in the ACP summer program.

*Call for Participation and Selection.* ACP did their traditional wide-spread mailing to thousands of scientists worldwide, and made publicly-advertised announcements.
regarding their summer program (and this workshop in particular, which was on the publicized agenda for the summer program). To augment these announcements, the Organizing Committee made a direct mailing to over two hundred scientists active in molecular biology, computer science, or mathematics (or interdisciplinary research among these three disciplines), encouraging them either to apply or to pass information about the workshop along to other potential participants. Participants were selected as described in the first workshop report.

Organization. Two formal talks were scheduled each day; the remainder of the time was devoted to small discussion groups, initiation of collaborations, research, and informal presentations.

Computational facilities. We set up a small, temporary workstation network for use by our workshop, including:

- Sparc20 (Sun Microsystems, Denver, CO),
- Sparc10 (Sun Microsystems, Los Alamos, NM),
- Sparc10 (Los Alamos National Laboratory, Los Alamos, NM),
- Sparc5 (Sun Microsystems, Denver, CO),
- Sparc1 (Los Alamos National Laboratory, Los Alamos, NM),
- Apple Laserprinter (Los Alamos National Laboratory, Los Alamos, NM),
- and Internet connection (Aspen Smallworks, Aspen, CO).

System administration of this network was provided by M. Engle, a workshop participant and a systems programmer in the Theoretical Biology and Biophysics Group at Los Alamos National Laboratory. The workstations were connected to a dedicated ISDN line providing a 56K BPS link to the Internet.

Software and databases that were available included:

- standard UNIX utilities (file editors and management, etc.);
- the TROFF electronic typesetting suite;
- copies of the GenBank, PIR, SwissProt, and Entrez databases;
- a copy of D. Higgins' CLUSTAL for multiple sequence alignments;
- a copy of W. Pearson's FASTA for scanning databases;
- a copy of S. Henikoff's BLOCKS software;
- a copy of X. Huang's CAP sequence assembly software;
- a copy of C. Burks' GENFRAG sequence fragmentation software;
• a copy of S. Smith's GDE software suite;
• several windowing systems, including OpenWindows 3.0 and X-11 R5;
• the xfig drawing program;
• and compilers, including FORTRAN, Pascal, and C.

The computational resources served as an advanced 'electronic blackboard' where new ideas and data sets could be laid out and modified more rapidly than would otherwise be possible. As such, it was a relatively unique aspect of this workshop compared to others (though several 'genome' workshops now routinely provide workstations for demonstrating software already developed, they are rarely used as a focus of initiating and carrying through on new ideas during the meeting). In addition, the facility allowed participants to communicate via e-mail with their home institution and other colleagues, an important consideration when one is leaving one's workplace for up to three weeks.

The majority of applicants made a substantial commitment of time, with the minimum stay being one week and a number attending for the full three weeks. The following is a list of those who participated in the workshop. This group represented an excellent cross-section of the disciplines and expertise the workshop topics drew on, and included two graduate students and four post-doctoral fellows. Sixty percent of the participants had not previously attended any of the workshops in our series.

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Formal Background</th>
<th>Research Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pankaj Agarwal</td>
<td>Washington U.</td>
<td>computer science</td>
<td>DNA sequence patterns</td>
</tr>
<tr>
<td>George Bell</td>
<td>LANL</td>
<td>physics, biophysics</td>
<td>DNA sequence patterns</td>
</tr>
<tr>
<td>William Bruno</td>
<td>LANL</td>
<td>physics, biophysics</td>
<td>sequence similarity</td>
</tr>
<tr>
<td>Christian Burks</td>
<td>LANL</td>
<td>databases, software, molecular biology</td>
<td>DNA sequence patterns, DNA sequence assembly</td>
</tr>
<tr>
<td>Francisco De La Vega</td>
<td>CINVESTAV-IPN (Mexico)</td>
<td>genetics, molecular biology</td>
<td>DNA sequence patterns, DNA sequence assembly</td>
</tr>
<tr>
<td>Michael Engle</td>
<td>LANL</td>
<td>computer science</td>
<td>software engineering, sequence similarity</td>
</tr>
<tr>
<td>Rob Farber</td>
<td>LANL</td>
<td>computer science</td>
<td>protein sequence patterns, protein folding</td>
</tr>
<tr>
<td>Richard Fye</td>
<td>Sandia Nat'l Lab</td>
<td>computer science</td>
<td>DNA structure &amp; dynamics</td>
</tr>
<tr>
<td>David Haussler</td>
<td>UC Santa Cruz</td>
<td>computer science</td>
<td>DNA sequence patterns, machine learning methods</td>
</tr>
<tr>
<td>Gerald Hertz</td>
<td>U. Colorado</td>
<td>molecular biology</td>
<td>DNA sequence patterns, DNA sequence similarity</td>
</tr>
<tr>
<td>Name</td>
<td>Institution</td>
<td>Formal Background</td>
<td>Research Focus</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------------------</td>
<td>----------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Deborah Joseph</td>
<td>U. Wisconsin</td>
<td>computer science</td>
<td>RNA secondary structure, DNA sequence similarity, DNA sequence assembly</td>
</tr>
<tr>
<td>Leonid Kruglyak</td>
<td>Whitehead, MIT</td>
<td>physics, biophysics</td>
<td>genetic mapping, DNA sequencing</td>
</tr>
<tr>
<td>Semyon Kruglyak</td>
<td>Cornell</td>
<td>computer science, mathematics</td>
<td>operations research</td>
</tr>
<tr>
<td>Alan Lapedes</td>
<td>LANL</td>
<td>physics</td>
<td>DNA sequence patterns, protein folding</td>
</tr>
<tr>
<td>Webb Miller</td>
<td>Penn. State U.</td>
<td>computer science, software</td>
<td>sequence similarity, gene regulation</td>
</tr>
<tr>
<td>Gene Myers</td>
<td>U. Arizona</td>
<td>computer science, software</td>
<td>sequence similarity, DNA sequence patterns, DNA sequence assembly</td>
</tr>
<tr>
<td>Bill Pearson</td>
<td>U. Virginia</td>
<td>biochemistry, software</td>
<td>sequence similarity</td>
</tr>
<tr>
<td>Sh'muel Pietrokovski</td>
<td>Fred Hutchinson Cancer Research Center</td>
<td>molecular biology</td>
<td>sequence similarity, DNA sequence patterns</td>
</tr>
<tr>
<td>Margaret Pullen</td>
<td>Evergreen Scientific Applications, Inc.</td>
<td>physics, mathematics</td>
<td>gene regulation</td>
</tr>
<tr>
<td>David Rabson</td>
<td>LANL</td>
<td>physics</td>
<td>protein folding</td>
</tr>
<tr>
<td>Hershel Safer</td>
<td>Genome Therapeutics, Inc.</td>
<td>computer science, software</td>
<td>DNA sequence assembly</td>
</tr>
<tr>
<td>Michael Smith</td>
<td>PRI/Dyncorp and NCI</td>
<td>molecular genetics</td>
<td>DNA mapping, DNA sequence patterns, DNA sequence polymorphism</td>
</tr>
<tr>
<td>Victor Solovyev</td>
<td>Baylor</td>
<td>mathematics, software</td>
<td>DNA sequence patterns</td>
</tr>
<tr>
<td>Paul Stolorz</td>
<td>Jet Propulsion Lab</td>
<td>physics, biology</td>
<td>RNA folding, protein folding</td>
</tr>
<tr>
<td>Gary Stormo</td>
<td>U. Colorado</td>
<td>molecular biology, software</td>
<td>DNA sequence patterns, RNA folding</td>
</tr>
<tr>
<td>Denis Thieffry</td>
<td>U. Libre de Buxelle (Belgium)</td>
<td>biology</td>
<td>DNA sequence patterns, gene regulation</td>
</tr>
<tr>
<td>Erik Wallin</td>
<td>Stockholm U. (Sweden)</td>
<td>physics</td>
<td>protein sequence patterns, protein folding</td>
</tr>
</tbody>
</table>
The table below is a list of the formally scheduled talks given during the workshop (there were additional, informal presentations and discussions). The focus ranged from work already completed, to current problems, to future strategies to address them. Several of the talks were tutorial in nature.

<table>
<thead>
<tr>
<th>Name</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Pearson</td>
<td>Overview of sequence similarity search and comparison of algorithms</td>
</tr>
<tr>
<td>Bill Pearson</td>
<td>Overview of sequence-based phylogenetic tree methods</td>
</tr>
<tr>
<td>Bill Bruno</td>
<td>Profile optimization and phylogenetic trees</td>
</tr>
<tr>
<td>Deborah Joseph</td>
<td>16S rRNA alignments and trees in soil microbial communities</td>
</tr>
<tr>
<td>Mike Smith</td>
<td>Microsatellites and genotyping</td>
</tr>
<tr>
<td>Gerry Hertz</td>
<td>Computational algorithms for identifying consensus patterns</td>
</tr>
<tr>
<td>George Bell</td>
<td>Evolution of microsatellite repeat distributions</td>
</tr>
<tr>
<td>Pankaj Agarwal</td>
<td>Evolution of repetitive DNA</td>
</tr>
<tr>
<td>David Haussler</td>
<td>Hidden Markov models for multiple protein sequence alignment</td>
</tr>
<tr>
<td>Hershel Safer</td>
<td>Genome sequencing: practical issues</td>
</tr>
<tr>
<td>Victor Solovyev</td>
<td>Tuple-based Bayesian methods for gene finding</td>
</tr>
<tr>
<td>Rob Farber</td>
<td>Neural network representations of protein potentials</td>
</tr>
<tr>
<td>Francisco de la Vega</td>
<td>Codon usage in lambda bacteriophage</td>
</tr>
<tr>
<td>Sh’muel Pietrokovski</td>
<td>Inteins: self-splicing proteins</td>
</tr>
<tr>
<td>Denis Thieffry</td>
<td>Modeling relationships between regulatory regions using logical expressions</td>
</tr>
<tr>
<td>Richard Fye</td>
<td>A model for describing DNA melting</td>
</tr>
<tr>
<td>Gene Myers</td>
<td>Fragment assembly: simplification of layout determination</td>
</tr>
<tr>
<td>Webb Miller</td>
<td>A globin gene server: multiple alignments and gene regulation</td>
</tr>
<tr>
<td>Christian Burks</td>
<td>Layouts for clone end-sequence sampling data</td>
</tr>
<tr>
<td>Leonid Kruglyak</td>
<td>De-mystifying linkage analysis: application of hidden Markov models</td>
</tr>
<tr>
<td>Erik Wallin</td>
<td>Predicting membrane spanning regions of membrane proteins</td>
</tr>
</tbody>
</table>

In addition, two of the participants (F. de la Vega and W. Pearson) participated in the GNA-VSNS Biocomputing Course, available over the Internet, during the course of the workshop and through access provided by our small workshop network. Finally, W. Pearson delivered a public lecture on the Human Genome Project during the course of the workshop.

We know of fifty-five papers that explicitly acknowledge the Aspen Center for Physics for work initiated or advanced during attendance at this workshop series. Participants have also informed us of about fifteen additional papers that were influ-

18 January 1996

ACP Workshop Report, May-June 1995
enced by the workshops, though without explicit acknowledgment. We anticipate that the number of such examples will continue to increase as ongoing work is published.

In addition, the Organizing Committee wrote this summary.\footnote{5}

We are grateful to the Office of Health and Environmental Research of the Department of Energy for the funds they provided for our workshop, to the NSF for their support of the ACP summer program in general (#PHY-9104428), and to Sun Microsystems, Aspen Smallworks, and Los Alamos National Laboratory for their provision of hardware, software, and communications lines. We also acknowledge the able administrative support from the ACP staff.

**General Citations:**


**Citations Explicitly Acknowledging ACP:**


18 January 1996

ACP Workshop Report, May-June 1995
Additional Papers Influenced by Participation in the ACP Workshops:


