IDENTIFYING FEATURES IN BIOLOGICAL SEQUENCES:

FIFTH WORKSHOP REPORT

(ASPEN CENTER FOR PHYSICS, MAY 30 - JUNE 19, 1994)

Eugene Myers¹, Christian Burks³, and Gary Stormo⁴

TR 94-36
Dept. of Computer Science
University of Arizona
Tucson, AZ 85721
December 19, 1994

Abstract: This report is for the fifth of an annual series of workshops held at the Aspen Center for Physics concentrating on the identification of features in DNA sequence, and more broadly on related topics in computational molecular biology. Over the last five years the workshop has been cited as supporting and/or inspiring over 40 papers, has provided training and inspiration to Ph.D students and post-doctoral fellows early in their careers, and provided senior scientists the unique opportunity to seriously engage in interdisciplinary collaborations over the 3 weeks of the workshop.

¹To whom correspondence should be addressed: telephone, 602-621-6612 fax, 602-621-4246 email: gene@cs.arizona.edu
²Dep’t of Computer Science University of Arizona Tucson, AZ 85721
³Theoretical Biology and Biophysics Group Los Alamos National Laboratory Los Alamos, NM 87545
⁴Dep’t of Molecular, Cellular, and Developmental Biology University of Colorado Boulder, CO 80309
Summary

The Aspen Center for Physics (ACP), in Aspen, Colorado, sponsored a three-week workshop from May 30 to June 19, 1994, with 23 scientists participating, 13 for their first time. The workshop, entitled Identifying Features in Biological Sequences was the fifth (IF-V) in a yearly series hosted by ACP on this topic. The previous workshops (RG-I, RG-II, RG-III, and IF-IV) occurred in years 1990 through 1993.

The workshop focused on discussion of current needs and future strategies for developing the ability to identify and predict 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 earlier in the report on RG-I. Of particular interest over the past year has been the maturation of previously described\(^5,6,9,14\) as well as the emergence of several new\(^4,7,24,29,45–49\) approaches to predicting the presence and location of genes (or at least protein-coding exons).

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 as genes) exclusively; for this reason, RG-I and RG-II have 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
- with scientists from disparate disciplines
- and a meeting 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.
Administrative Details


Location of the workshop. The ACP 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 previously.

Organizing Committee. E. Myers, C. Burks, and G. Stormo constituted the Organizing Committee, with E. Myers 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. Funding was provided through the standing NSF grant (NSF-86-06266) to ACP for its summer program, and from NSF grant BIR-9406201 specifically awarded to this year’s workshop. These funds were used to partially offset the participants’ costs in attending the workshop and 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 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. Selection of participants was made as described in the RG-I report.

Workshop 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 three Sun workstations (1 LX, 1 Classis, and 1 Sparc 1) and three additional Sun X-terminals. All but one Sun workstation were loaned to us for the duration of the workshop by Sun Microsystems Denver office to whom we are very grateful. The Sparc 1 and a post-script capable laser writer were loaned to us by Los Alamos National Laboratory along with an additional gigabyte of disk to accompany a gigabyte loaned by Sun. 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 and X-terminals were configured in a network that included a direct "slip-link" to the Internet (with the required hardware, installation, and systems consulting very purchased from the Colorado SuperNet), and modems for dialing remote computers and allowing remote computers to dial in.

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 dna sequencing software;
a copy of C. Burks' GENFRAG dna sequencing software;
a copy of S. Smith's GDE software suite;
several windowing systems, including OpenWindows 2.0, X-11 R5, and SunView;
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 email with their home institution and other colleagues, an important consideration when one is leaving their workplace for up to three weeks.

Participants

The majority of applicants made a substantial commitment of time, with the minimum stay being one week and many attending for the full three weeks. The following is a list of those who participated in the workshop (unfortunately, there were more applicants than there were slots available). This group represented an excellent cross-section of the disciplines and expertises the workshop drew on, and included four graduate students and four post-doctoral fellows. One-half of the participants in IF-V had not attended a previous Aspen workshop.

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Formal Background</th>
<th>Research Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tim Bailey</td>
<td>UCSD</td>
<td>Computer Science</td>
<td>pattern-learning, biocomputing overview, intron structure, experimental biology</td>
</tr>
<tr>
<td>George Bell</td>
<td>LANL</td>
<td>Physics, Biophysics</td>
<td></td>
</tr>
<tr>
<td>Christian Burks</td>
<td>LANL</td>
<td>Databases, Software, Biology</td>
<td>pattern matching, data structures, high speed searches</td>
</tr>
<tr>
<td>Sean Eddy</td>
<td>MRC-LMB</td>
<td>Biology, Software</td>
<td>sequence alignment, HMM models</td>
</tr>
<tr>
<td>Michael Engle</td>
<td>LANL</td>
<td>Computer Science</td>
<td>software engineering, database search algorithms</td>
</tr>
<tr>
<td>Rob Farber</td>
<td>LANL</td>
<td>Computer Science</td>
<td>neural nets, protein folding</td>
</tr>
<tr>
<td>D'vorah Graeser</td>
<td>Wash U.</td>
<td>Biology</td>
<td>hybridization technology, physical mapping</td>
</tr>
<tr>
<td>George Hartzell</td>
<td>Berkeley</td>
<td>Computer Science, Biology, Software</td>
<td>dna sequencing, physical mapping</td>
</tr>
<tr>
<td>Jim Holloway</td>
<td>Oregon St.</td>
<td>Biology, Computer Science</td>
<td>sequence comparison, genetic mapping</td>
</tr>
<tr>
<td>Speaker</td>
<td>Institution</td>
<td>Department, Field</td>
<td>Topic</td>
</tr>
<tr>
<td>------------------------</td>
<td>--------------</td>
<td>--------------------------------------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>Paul Horton</td>
<td>Berkeley</td>
<td>Computer Science</td>
<td>multi-alignment</td>
</tr>
<tr>
<td>Gordon Hutchinson</td>
<td>U. British Columbia</td>
<td>Medicine, Biology</td>
<td>gene recognition, repetitive elements</td>
</tr>
<tr>
<td>Toni Kazic</td>
<td>Washington U.</td>
<td>Biology, Computer Science</td>
<td>genome informatics, metabolic pathways</td>
</tr>
<tr>
<td>Jim Knight</td>
<td>U. Arizona</td>
<td>Computer Science</td>
<td>pattern matching</td>
</tr>
<tr>
<td>Leonid Kruglyak</td>
<td>Whitehead (MIT)</td>
<td>Physics, Software</td>
<td>physical mapping</td>
</tr>
<tr>
<td>Alan Lapedes</td>
<td>LANL</td>
<td>Physics, Mathematics</td>
<td>neural nets, protein folding</td>
</tr>
<tr>
<td>Richard Lathrop</td>
<td>MIT AI</td>
<td>Computer Science</td>
<td>pattern learning &amp; matching, protein folding</td>
</tr>
<tr>
<td>Suzanna Lewis</td>
<td>LBL</td>
<td>Biology, Software, Computer Science</td>
<td>physical mapping databases</td>
</tr>
<tr>
<td>Catherine Macken</td>
<td>LANL</td>
<td>Biology, Software</td>
<td>gene recognition</td>
</tr>
<tr>
<td>Gene Myers</td>
<td>U. Arizona</td>
<td>Computer Science, Software</td>
<td>high speed searches, pattern languages</td>
</tr>
<tr>
<td>Bill Pearson</td>
<td>U. Virginia</td>
<td>Biochemistry, Software</td>
<td>scoring schemes, high speed searches</td>
</tr>
<tr>
<td>Victor Solovyev</td>
<td>Baylor</td>
<td>Biology, Software</td>
<td>sequence analysis software</td>
</tr>
<tr>
<td>Gary Stormo</td>
<td>U. Colorado</td>
<td>Biology, Software</td>
<td>predicting functional regions, pattern-learning, sequence signals</td>
</tr>
<tr>
<td>Michael Storrie-Lombardi</td>
<td>Cambridge</td>
<td>Astronomy, Physics</td>
<td>protein folding</td>
</tr>
</tbody>
</table>

Presentations made

The table below is a list of the formally scheduled talks given during the workshop (there were a number of more informal presentations/discussions). The focus was both on work already accomplished and on current problems and future strategies to address them. Several of the talks were tutorial in nature.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>George Bell</td>
<td>A taxonomy of repetitive DNA elements</td>
</tr>
<tr>
<td>Gordon Hutchinson</td>
<td>A taxonomy of Alus</td>
</tr>
</tbody>
</table>
Workshop citations.

We know of over 40 papers\textsuperscript{10–51} that explicitly acknowledge the Aspen Center for Physics on work initiated or advanced during attendance at RG-I, RG-II, RG-III, IF-IV, or IF-V. In addition, several other pieces of work\textsuperscript{52–57} were influenced by the workshops (but do not cite it directly). 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 report\textsuperscript{8}.
Acknowledgements.

We are extremely grateful to ACP for sponsoring this workshop, and to the Center for Human Genome Studies and the Theoretical Division at Los Alamos National Laboratory, for providing funds for this workshop. In addition, we are indebted to Colorado SuperNet, Inc. for donating the hardware and staff support for establishing and maintaining the slip-link connection to the Internet during the workshop; and to Sun Microsystems for donating a Sun workstation server during the workshop. The Theoretical Biology and Biophysics Group at Los Alamos National Laboratory loaned several Sun workstations and peripheral hardware to us; in addition, M. Engle and F. Martinez in that group provided systems planning and support for the workshop’s workstation network. We also acknowledge the able administrative support from the ACP staff.

References.


Papers That Cite The ACP Workshops:


**Papers That Were Influenced By Participation At The ACP Workshops:**


