Automated sequence reading and analysis

John West

Bio-Rad Laboratories, 3316 Spring Garden Street, Philadelphia, PA 19104, USA

Received August 17, 1987; Revised and Accepted December 31, 1987

ABSTRACT

We report on a system developed by Bio-Rad Laboratories, Inc. which combines automated reading of DNA sequencing autoradiograms with comprehensive software for shotgun overlapping of the readings, analysis of the sequences derived and searching of databases. Reading is accomplished using a high speed optical scanner and pattern recognition software operating on a personal computer. Overlapping, analysis and database searching software each incorporate significant advances over prior systems.

INTRODUCTION

Massive sequencing projects, such as the Human Genome Initiative, now loom on the horizon. Equally important are numerous intermediate-size sequencing projects either underway or in the planning stages. These projects call for automation of gel reading, consensus sequence assembly and interpretation. We consider here a system which addresses all of these needs in an integrated manner.

MATERIALS AND METHODS

Automatic Film Reading

In a landmark paper, Elder, Green and Southern demonstrated that automation of sequence gel reading is possible (1). Using a linear array detector, they were able to obtain digital images from autoradiograms, and by analyzing the data with pattern recognition algorithms, were able to assign a
sequence. They reported 99% accuracy. In a collaborative effort with the authors of this paper, Bio-Rad has further advanced this system.

A new scanner has been developed for digitizing sequencing autoradiograms. It is based on a 1,728 element CCD array detector (Texas Instruments). The film to be digitized is supported on a flat moving platen and is viewed by the camera in transmission. This scanner has been interfaced to a microcomputer (IBM AT Model 339) using DMA channels for both control and data taking. This means that scanner speed is no longer limited by software execution time and an image can be digitized with very high resolution in as little as 30 seconds. The circuitry allows completely arbitrary, smoothly variable motion of the sample. Thus, data points are typically taken with a spacing of 50 microns at the top of the film and below that the sample is continuously accelerated. This is used to match the scanning speed to the resolution required, and reduces the number of data points to be processed. The motion profile, determined by software, can be easily adapted to suit the band spacing profile of gradient or wedge spacer gels.

The software has been developed using the Xenix operating system. Xenix is a commercial implementation of the UNIX operating system widely used for scientific work. It allows access to the full 16 Megabyte address space of the microprocessor in the AT computer and provides both multi-user and multi-tasking operation. It also allows wider collaboration with other researchers using scientific workstations and mini computers. All programs have been written in "C", with key routines converted to assembly language for higher performance.

Pattern recognition proceeds in several steps. The first major step recognizes lanes, corrects for lane wander and band smiling, and derives a one dimensional densitometric function for each lane. The new algorithms developed for this step work as follows:
Raw data, output by the scanner, is first corrected for illumination non-uniformity and detector dark current. This results in a two dimensional array of optical density values. Lanes are recognized in five steps:

a) recognizing that the bands constitute high frequencies in a vertical spatial Fourier transform, they are extracted by a zero phase high pass digital filter:

\[
F(i) = O(i) - \frac{1}{2N+1} \sum_{i-N}^{i+N} O(i)
\]

b) Comparison with a noise threshold converts the result to a binary image.

c) Contiguous "active" pixels in the binary image are located and represented in dynamic data structures as "objects".

d) These objects are pruned and the list of them is then filtered against geometric criteria to eliminate artifacts like gel bubbles and stray radioactive specks. (See Figure 1.)

e) Near vertical chains of objects are recognized and identified as lanes.

Using the locations of the lanes in the image and the shape of the objects (bands), the program then performs a two to one dimensional compression resulting in a densitometric trace for each lane. Corrections for smiling and differential registration of the lanes are performed at this point. Median filtering further suppresses specks and voids in the original image.

Once a densitometric trace has been determined for each lane, sequence interpretation proceeds, using algorithms similar to those of Elder, et al.

Results are displayed on a high resolution color monitor. Both the assigned sequence and the digitized image are presented together, to aid verification. Details of the verification process will be presented elsewhere.
Figure 1:
(a) Section of a digitized and processed autoradiogram. Bands are outlined by the software as a step in the object-lane finding process.
(b) Presentation of the digitized image along with the assigned sequence, ready for operator review.
Shotgun Overlapping Software

Automation of film reading may lead to higher volume sequencing which, in turn, will put pressure on shotgun overlapping software. The system has been based on the DBUTIL and DBAUTO programs of Staden (2). The original Fortran programs have been rewritten in "C", but their key features have been retained. These include representation of contigs as structures of registered and edited gel readings while the archival gel sequences are kept untouched. Whole groups of gel readings can be processed automatically into contigs.

Beginning with this base, the following improvements have been made: when a new gel reading is being merged into a contig, the system performs a rigorous Needleman-Wunch alignment to determine where padding characters are required (3). The limited computer capabilities available when the original programs were written did not allow this thoroughness.

Tools have also been incorporated to automate user resolution of problems which the automatic algorithms were not able to handle. When manually entering a gel, the user can now ask for the best alignment the automatic program was able to achieve and edit, or reject, the registration from there. All editing, either of a new gel reading, a reading already in a contig, or of all gels at a particular contig position, can now be performed using a mouse. Display of the overlaps and consensus are automatically recalculated after each edit.

The user interface presents menus on the console and data on a graphics screen, so users no longer need lose their data off screen when menus are presented. A consistent menu-based user interface has been adopted which dramatically simplifies learning and operation of all programs.

Tools are also available to break a contig in two, or to dissolve it entirely, if errors in overlapping are found. The user can also back up to the contig structure which existed at a prior point, without losing gel readings which have been added to the project in the interim. To help a user
refer back to their films to resolve disagreements, the programs can also report a gel reading's sequence in its original sense. This is helpful if a gel sequence has been reversed and complimented in the process of entry into a contig.

The perspective of the shotgun programs has also been broadened to incorporate all gel readings in a project. The user is no longer burdened with creating, naming and keeping track of gel reading files and gel batch files. New readings can be submitted directly to a shotgun project on entry. The software tracks and can report on the progress of each gel reading through all stages of processing. Specific menus are available to recover gel readings which have failed restriction, or vector screening, or automatic entry.

Project management tools have also been added. The system can report on the state of gel sequence processing, the state of sequencing (number and size of contigs, extent of overlapping, etc.) and on the state of the consensus quality. The programs also automatically compile a history record showing the number of gel readings, contigs, etc. versus time and date.

Lastly, the shotgun handling process has been linked to all of the programs in the system. Any program can directly refer to shotgun project gel readings and contig consensus sequences without need for writing this data to user files as an intermediate step. The automatic film reading program can submit its results directly to a shotgun project for processing.

**Sequence Analysis Software**

The DNA sequence analysis software is based on the ANALYSEQ programs from Staden. The programs have been rewritten in "C", with assembly language used to improve performance, quite dramatically in some cases. The graphical presentation of results has been extended to one million pixel resolution, including mouse interaction for pointing to and measuring features, zooming in, etc. To this analytical base has been added programs for DNA and protein
sequence alignment, based on the algorithm of Needleman and Wunsch. This algorithm has been extended by implementation of a sophisticated gap penalty.

$$P = A_0 + A_1 N_g + A_2 F (N_g \text{MOD} 3)$$

The first term provides a constant penalty, independent of gap length, and the second term is linearly dependent on gap length. The last term provides a penalty for gaps which constitute a frame shift, useful for alignment of sequences known to code for proteins.

A protein sequence analysis program has also been developed, incorporating reverse translations, hydrophobicity and secondary structure analysis. A new algorithm has been included which plots the geometric average over a window of the reverse translation degeneracy of a protein sequence versus position along a sequence. This can help in identifying favorable locations for development of synthetic probes.

**Database Searching**

The system includes the complete Genbank and PIR databases on disk. Any entry in either database can be referred to directly by locus name in any of the programs in the system. In addition to homology and text matching sequence searches, the system includes a very rapid keyword search. To accomplish this every word from every field of every entry in each database has been extracted and organized into sorted tables which can be probed by a binary search algorithm. A search of the whole Genbank is generally complete in under ten seconds. This goes far beyond the index files provided on tape with the original database distribution. This approach can pick terms used anywhere in a sequence entry, not just keywords selected by the Genbank or PIR staffs. A single search can be used for locating entries by locus, accession number, author, keyword, organism, etc. Users do not have to be familiar with the structure of database entries to search the database, and the response is so fast that interactive searching becomes practical.
In addition to keyword searching, the software can find similarities between a query sequence and any portion of Gen Bank using the Needleman-Wunsch algorithm. The query can be as large as 10,000 bases. The multi-tasking feature of the software can place such searches into the background and let the user perform up to two other tasks concurrently.

RESULTS AND DISCUSSION

We have analyzed several types of films with this system. One film has 177 bases, the rest being off the top of the film. The machine achieved 100% accuracy over this range, in spite of the presence of two mild compressions. Algorithms in the software allow it to recognize the variation in band spacing which is characteristic of compressions. Another film was from a 100 cm long gel. This was scanned in three overlapping segments and the readings were automatically overlapped using the shotgun handling software. The overlapped result showed one error in the first 240 bases, this due to anomalous spacing on the gel itself.

For comparison, we have asked five experienced researcher to read a film of a known sequence (M13). Readers were given as long as they wanted and generally took 30-60 minutes (compared with 5 minutes for the machine). Readers stopped on average after 250 bases and achieved a 3.4% error rate. The most common source of these errors was misinterpretation of the number of bases in a run of consecutive bases of the same assignment (e.g. ccc vs cccc). As projects grow in size and the number of bands to be read reaches into the hundreds of thousands and even millions, we suspect that, if human readers are used, the error rate will increase from what has been found under these test conditions.

Comparison of the Bio-Rad film reading system with alternative automation approaches (4) highlights some advantages. First, the solid state scanning system offers dramatically lower costs, faster scanning and
higher reliability than laser scanners. Secondly, since electrophoresis and measurement are decoupled in this system, expensive measurement equipment is not tied up during long electrophoresis runs. This markedly improves throughout per capital equipment expenditure and consequently reduces cost per base. Since conventional chemistry is used and a complete image is acquired and analyzed, well documented artifacts are involved. They are visible in the image and can be corrected in the software. When discrepancies are found through the shotgun overlapping process, they can be resolved by referring back to the film or the on-line digital image. None of this is possible when using detection on-the-fly from the gel. A scanner for sequencing autoradiograms can also be a flexible laboratory tool, used for scanning restriction digests and standard one and two dimensional electrophoresis gels.

The system's shotgun overlapping software is moving toward support of massive sequencing projects. It can improve the quality of automatically determined alignments, broaden the scope of gel processing automation and provide for computer-assisted discrepancy resolution and project management. As sequencing projects grow, management of the large volume of gel reading data will certainly grow in importance. The system's analysis and database searching software also improve the level of automation and user productivity, both of which will be crucial as sequencing volume rises.

**CONCLUSION**

The system which has been reviewed here, is the only system of which we are aware that integrates automatic reading of sequencing gels with comprehensive shotgun overlapping, analysis and database searching software in a multi-user environment suitable for team sequencing.

The cost per base of this film reading system is dramatically lower than alternative approaches and its integration with a complete sequencing software system provides additional economy.
REFERENCES

1. Elder, J.K., Green, D.K. and Southern, E.M., Automatic Reading of DNA Sequencing Gel Autoradiographs Using a Large Format Digital Scanner, Nucl. Acids Res., 14, 417-424 (1986).

2. Staden, R., Sequence Data Handling by Computer, Nucl. Acids Res., 4, 4037-4051 (1977).

   Staden, R., Further Procedures for Sequence Analysis by Computer, Nucl. Acids Res., 5, 1013-1015 (1978).

   Staden, R., A Strategy of DNA Sequencing Employing Computer Programs, Nucl. Acids Res., 6, 2601-2610 (1979).

   Staden, R., A New Computer Method for Storage and Manipulation of DNA Gel Reading Data, Nucl. Acids Res., 8, 3673-3694 (1980a).

   Staden, R., A Computer Program to Search for tRNA Genes, Nucl. Acids Res., 8, 817-825 (1980).

   Staden, R., Automation of the Computer Handling of the Gel Reading Data Produced by the Shotgun Method of DNA Sequencing, Nucl. Acids Res., 10, 4731-4751 (1982).

   Staden, R., A Computer Program to Enter DNA Gel Reading Data into a Computer, Nucl. Acids Res., 12, 499-503 (1984a).

   Staden, R. and McLachlan, A.D., Codon Preference and its Use in Identifying Protein Coding Regions in Long DNA Sequences, Nucl. Acids Res., 10, 141-156 (1982).

   Staden, R., An Interactive Graphics Program for Comparing and Aligning Nucleic and Amino Acid Sequences, Nucl. Acids Res., 10, 2951-2961 (1982b).

   Staden, R., Measurements of the Effects that Coding for a Protein has on a DNA Sequence and Their Use for Finding Genes, Nucl. Acids Res., 12, 551-567 (1984a).

   Staden, R., Computer Methods to Locate Signals in Nucleic Acid Sequences, Nucl. Acids Res., 12, 505-519 (1984b).

   Staden, R., Graphic methods to Determine the Function of Nucleic Acid Sequences, Nucl. Acids Res., 12, 521-538 (1984c).

3. Needleman, S.B. and Wunsch, C.D., A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, J. Mol. Bio, 48, 443-453 (1970).

4. Smith, L. M. et al., "Flourescence detection in automated DNA sequence analysis," Nature (London) 321, 674 (1986).

Xenix is a trademark of Microsoft Corporation.
Unix is a trademark of A T & T Bell Laboratories.