SANDIA REPORT

SAND96–2579 • UC–705
Unlimited Release
Printed November 1996

Fast DNA Sequence Alignment Using Optical Computing

Mark L. Yee, David C. Craft

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550
for the United States Department of Energy
under Contract DE-AC04-94AL85000

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
Office of Scientific and Technical Information
PO Box 62
Oak Ridge, TN 37831

Prices available from (615) 576-8401, FTS 626-8401

Available to the public from
National Technical Information Service
US Department of Commerce
5285 Port Royal Rd
Springfield, VA 22161

NTIS price codes
Printed copy: A03
Microfiche copy: A01

SAND96-2579
Unlimited Release
Printed November 1996

Distribution
Category UC-705

FAST DNA SEQUENCE ALIGNMENT USING OPTICAL COMPUTING

Mark L. Yee and David C. Craft
Exploratory Image Processing Systems II Department
Sandia National Laboratories
Albuquerque, NM 87185-0843

Abstract

Alignment of DNA sequences is a necessary step prior to comparison of sequence data. High-speed alignment is needed due to the large size of DNA databases. Correlation, a standard pattern recognition technique, can be used to perform alignment. Correlation can be performed rapidly using optical techniques. Thus, optical correlation offers the potential for high-speed processing of DNA sequence data. This report describes research efforts to apply one-dimensional acousto-optical correlation methods to the problem of DNA sequence alignment. Experimental results are presented.

Figures

Fast DNA Sequence Alignment Using Optical Computing

Introduction

Researchers often wish to compare different sequences of deoxyribonucleic acid (DNA) with one another to determine any similarities or differences between them. Sequences need to be aligned with one another prior to comparison. Alignment can be accomplished by using a correlation process, a standard technique used in pattern recognition. Correlation can be performed rapidly using optical techniques. Thus, optical correlation offers the potential for high-speed processing of DNA sequence data.

This report describes research efforts to apply one-dimensional optical correlation methods to the problem of DNA sequence alignment.

The DNA Sequence Alignment Problem

Part of the Human Genome Project involves the compilation of a massive database containing DNA sequence information obtained by molecular biology researchers throughout the world. A DNA sequence consists of a listing of the order of the four nucleic acid bases which comprise a particular strand of DNA. The nucleic acid bases are represented in the database by the letters G (guanine), C (cytosine), A (adenine), and T (thymine). Thus a typical sequence might appear as AAGCT. Researchers often wish to compare sequences with one another to determine any similarities or differences between them. For instance, a difference of one base in a particular sequence might indicate a mutation has taken place. Comparison of DNA sequences requires the sequences to be *aligned* with one another, as shown in Figure 1.

Figure 1. Alignment of Sequences

Sequence alignment is a simple matter when there are only a few bases per sequence, but DNA sequences of interest typically have hundreds or thousands of bases present. Large sequences will require long times to compute the correct alignment. Since alignment is only the first step of many in the sequence comparison problem, it is of interest to reduce the time required for alignment as much as possible. Existing alignment algorithms utilize dynamic programming methods to minimize a scoring function based on the matches between the sequences for various alignments. Since this is essentially an efficient global search, the combinatorics of large sequences will result in prohibitive computation times if serial processing is used. In response, researchers sometimes only align subsections of the sequences to save time. New computing solutions using systolic arrays and/or massively parallel computers have been proposed and tried.

Researchers realize that fast alignment methods will be invaluable to use of the massive amounts of data stored in genome databases in the future.

It is useful to underscore the difference between DNA sequence *analysis* and DNA sequence *alignment*. Sequence alignment is a subset of sequence analysis. As stated previously, sequence alignment is a first step towards the larger class of techniques used for seqeuence analysis. One example of this difference is the use of insertions and deletions of sections of DNA for comparison of sequences. A given sequence of DNA obtained from the laboratory may not be 100% correct in its content, small sections of the nucleic acid base sequence may be missing or added. A full analysis requires an algorithm that "inserts" or "deletes" small sections of the sequence during comparison. Sequence alignment by itself does not perform any insertion or deletion operations. A further distinction must be made from the term DNA *sequencing*, which is the operation of assembling several different sections, or "fragments" of DNA into a longer sequence. This is completely different from what is meant by sequence *analysis*.

The Application of Correlation to Sequence Alignment

Begin by treating the DNA sequence information as a one-dimensional signal. The alignment problem can then be solved by cross-correlating two signals, or sequences, against one another. Correlation yields a maximum value when two signals are identical, thus the correlation result can be used to indicate when the two signals (sequences) are aligned.

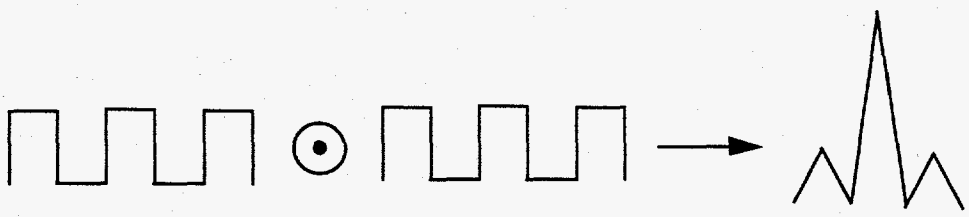

Figure 2. Correlation Result.

Application of correlation is thus facilitated by developing a proper binary code to represent each of the different nucleic acid bases (G,C,A,T), along with a code for unknown entries (X). The desirable properties of the code are that it result in unambiguous matches between the different bases, while minimizing the number of bits per base required. The need for unambiguous matches is obvious, while a minimum number of bits would minimize the computation time required.

A code which satisfies both constraints is the following :

A	1000
T	0001
G	0010
C	0100
X	1111

Table 1. Code for bases.

The unknown entry (X) is a "wild card" which will match any of the nucleic acid bases. The assumption is that any unknown entries could potentially be any one of the bases, thus it is considered as a match with any of them.

The correlation result from using such a code cannot be used directly. Instead, the result of interest is obtained by subsampling the correlation result by four. This is due to the fact that each "symbol" of interest is represented by a four-bit code. The correlation result for the "symbols" is thus represented by every fourth correlation value. This is shown by the following :

$$C(m) = \sum_{n} x(n)y(n+m) \qquad (1)$$

where C is the result of correlating signals x and y. The parameter m is a count of the relative "shift" between x and y. The four-bit symbols in both x and y will be aligned with each other only for shift values m that are divisible by four. Thus only every fourth value of $C(m)$ is useful.

An inverse code can also be employed :

A	0111
T	1110
G	1101
C	1011
X	1111

Table 2. Inverse code for bases.

When the code of Table 1 is used for one signal (x) and the code of Table 2 is used for the other signal (y), the correlation result is a minimum (instead of a maximum) at the point of highest match. Detecting a minimum value is easier to implement in the optical hardware. It is still true that only every fourth value of $C(m)$ is useful.

Actual DNA sequences were encoded according to these coding schemes. The sequences were then intentionally misaligned artificially, and put through a simulation of the correlation process. The amount of misalignment was detected accurately each time. Subsets of the DNA sequences were also successfully detected and located within the larger sequences by using the simulated correlation process.

Two-dimensional Optical Correlation

An experiment was performed to implement the DNA correlation scheme on an existing two-dimensional acousto-optical (AO) correlator. Details of the two-dimensional correlator are given elsewhere [1]. It is sufficient to understand that the two-dimensional correlator performs correlations between two-dimensional (not one-dimensional) signals. The code for the nucleic acid bases was modified to be implemented in a two-dimensional mode, where each nucleic acid base was represented by a four-pixel "square", with each pixel turned "on" or "off" according the the codes given in Tables 1 and 2. The original DNA sequences were then coded as separate two-dimensional images. Due to system constraints, one of the sequences was much shorter than the other.

There were no useful results obtained from the two-dimensional correlator. The correlator lacks sufficient illumination power to generate useful results with patterns that have few pixels which are turned "on", which is the case for the codes used for the nucleic acid bases. This effect has been noted in previous work involving the two-dimensional correlator, so results such as these were not unexpected. Use of the inverse code failed to alleviate the problem, as the low light levels made it impossible to detect even a minimum (rather than maximum) output. It is possible to compensate for the lack of illumination power by increasing the number of pixels (bits) used to encode each nucleic acid base, but the resulting increase in bandwidth would drastically reduce the computation speed of the correlator, making it impractical for this application.

The experiment demonstrated that it is not practical to apply the two-dimensional acousto-optical correlator (in its current form) to the problem of DNA sequence alignment.

One-dimensional Optical Correlation

One of the goals of this project was to design and construct a one-dimensional acousto-optical correlator optimized for the purpose of DNA sequence alignment. An optimal design would avoid the problems encountered using the existing two-dimensional correlator.

A conceptual layout of a one-dimensional acousto-optical correlator is shown in Figure 3.

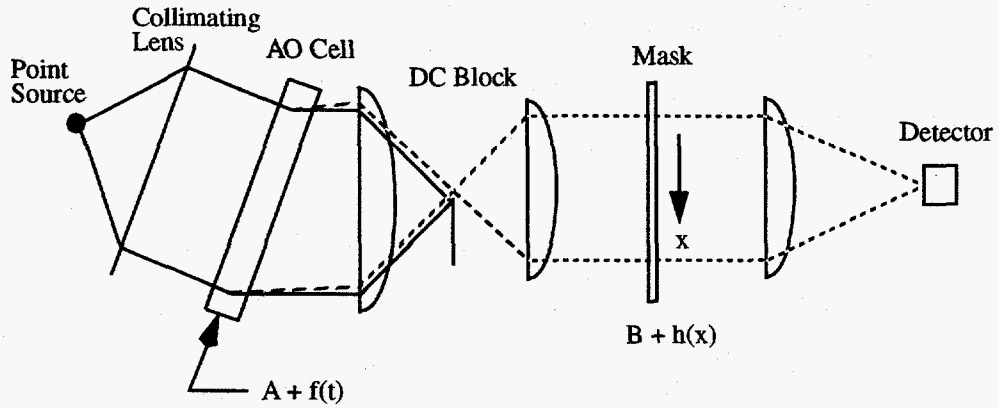

Figure 3. One-dimensional Correlator

The particular type of correlator shown in Figure 3 is a space-integrating correlator [1]. The input light is constant. The input light must be an approximate point source in order to focus the result on the single detector at the end. The input signal power $A + f(t)$ is input to the AO cell used in the Bragg diffraction regime. This causes the diffracted order to have the input signal imposed on it. The diffracted order only is passed and is incident on the mask. The mask has an intensity transmittance of $B + h(x)$. (It should be noted that a second AO cell was used instead of a fixed mask, thus enabling the signal h to be changed as needed.) The mask coordinates are opposite to the direction of propagation of the signal in the AO cell, as shown by the arrow at the mask in Figure 3. The transmitted intensity from the mask is then

$$[A + f(x - v_s t)][B + h(x)] = AB + Bf(x - v_s t) + Ah(x) + f(x - v_s t)h(x) \qquad (2)$$

The final lens then spatially integrates the intensity by focusing the light onto the single detector as shown, resulting in

$$ABL + B\int_L f(x - v_s t)dx + A\int_L h(x)dx + \int_L f(x - v_s t)h(x)dx \qquad (3)$$

where L is the length of the effective aperture of the AO cell. The first term is simply a constant, and the second and third terms will be zero if f and h are zero mean. (with the usual bias terms present if they are not). This leaves the last term, which is the desired correlation result

$$C_O(t) = \int_L f(x - v_s t)h(x)dx \qquad (4)$$

Equation (4) shows that the output of the space-integrating correlator is a function of time, and thus is read out sequentially from the detector. The integration length of L must be equal to the duration of the signals f and h. In this case, however, L is limited by the effective AO cell aperture. Thus the durations of f and h are limited by the system, as is the time-bandwidth product and the peak-to-sidelobe ratios. However, readout speed is limited only by the AO cell bandwidths and the readout speed of the detector.

It should be noted that the since a second AO cell is used instead of a fixed mask, the second term in the integral of Equation (4) becomes $h(x\text{-}v_s t)$ yielding $C_O(2t)$. Thus, the output from the AO correlator is the true correlation result subsampled by a factor of two. This is not a problem, since the coding scheme for the nucleic acid bases requires subsampling by a factor of four.

DNA Sequence Correlator Setup

The actual benchtop one-dimensional correlator system is described in three parts. The first part consists of the optics used to collimate, expand, and launch the input illumination from a laser diode, and is illustrated in Figure 4. All focal lengths are in mm. The main purpose of the launch optics is to create a collimated beam to fill the 40 mm aperture of the AO cells.

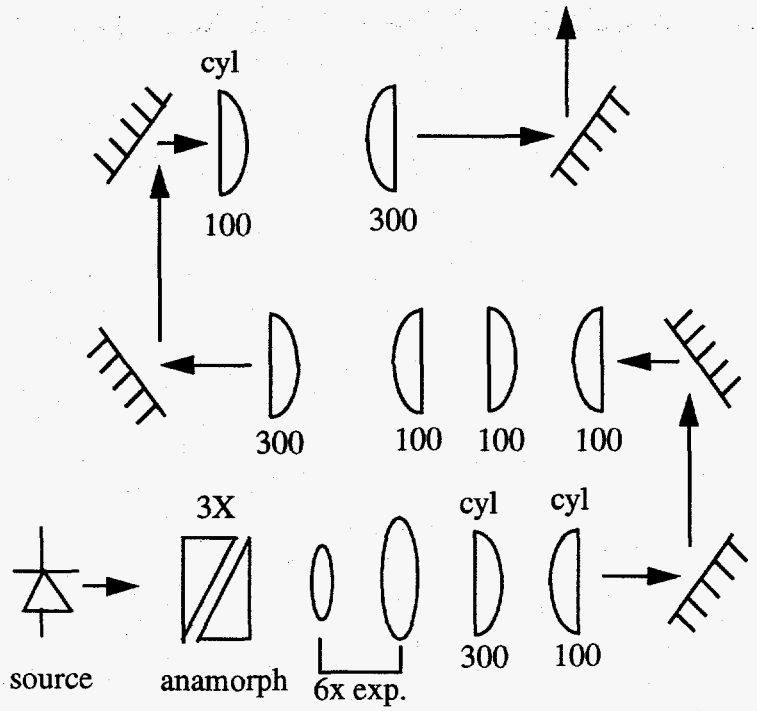

Figure 4. Launch Optics.

The 1.5 x 40 mm collimated light is then directed to the second part of the system consisting of the AO cells. The diffracted light from the first AO cell is imaged onto the second AO cell. This accomplishes the multiplication of the two input signals.

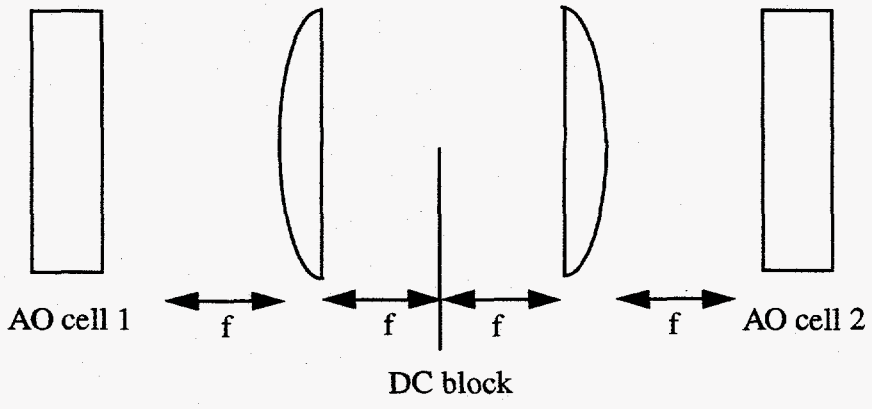

Figure 5. AO Cells with Transfer Optics

The diffracted light from the second AO cell is then focused onto a point detector to produce the correlation result.

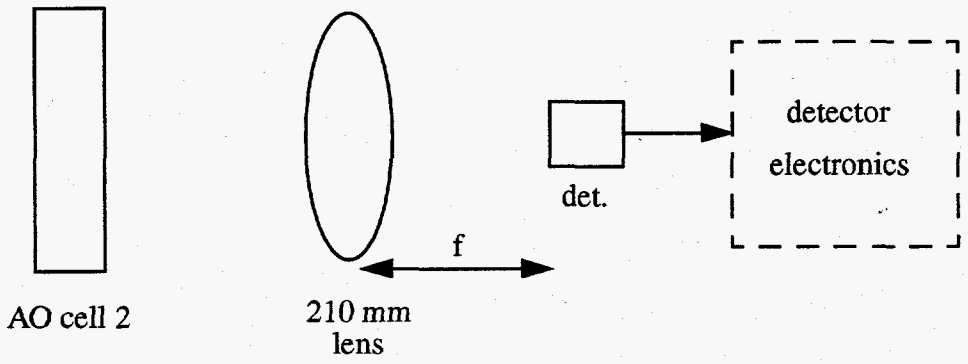

Figure 6. Detector

Experimental Results

The following correlation result was output from the optical correlator. Inputs were two square functions.

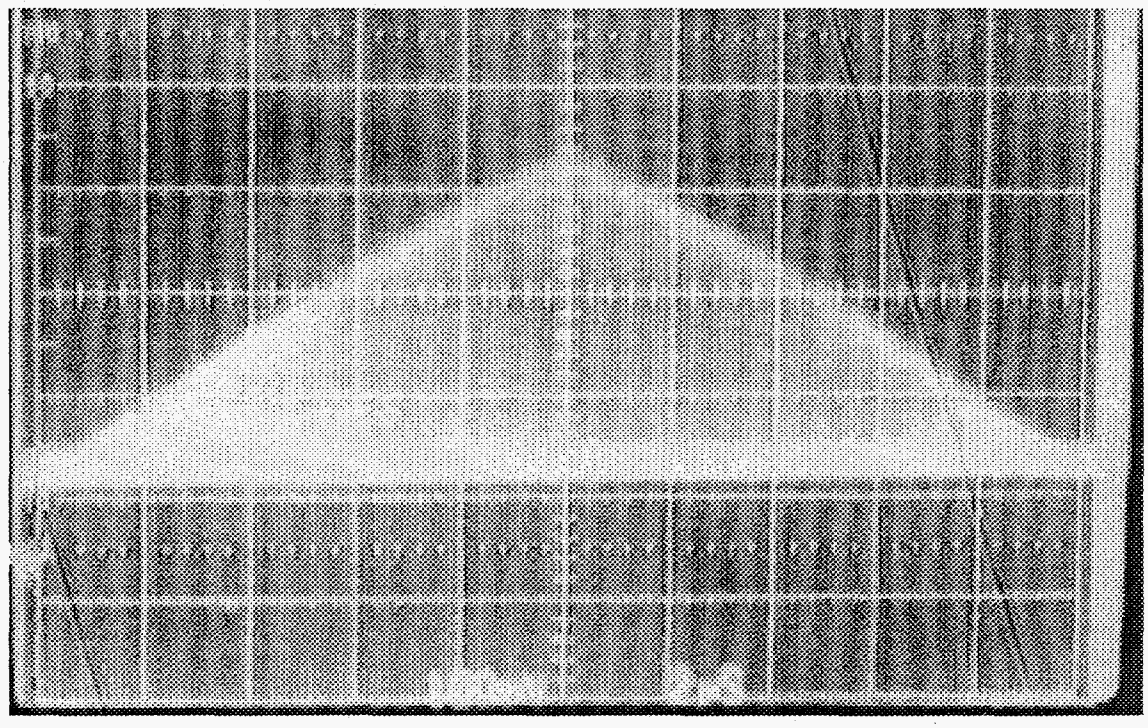

Figure 7. Correlation Output.

Figure 8 shows the same output on an expanded time scale. The period of the waveform is 80 nanoseconds, corresponding to a data rate of 12.5 MHz. This is the expected data rate for two 50 MHz bandwidth AO cells in a space-correlating configuration.

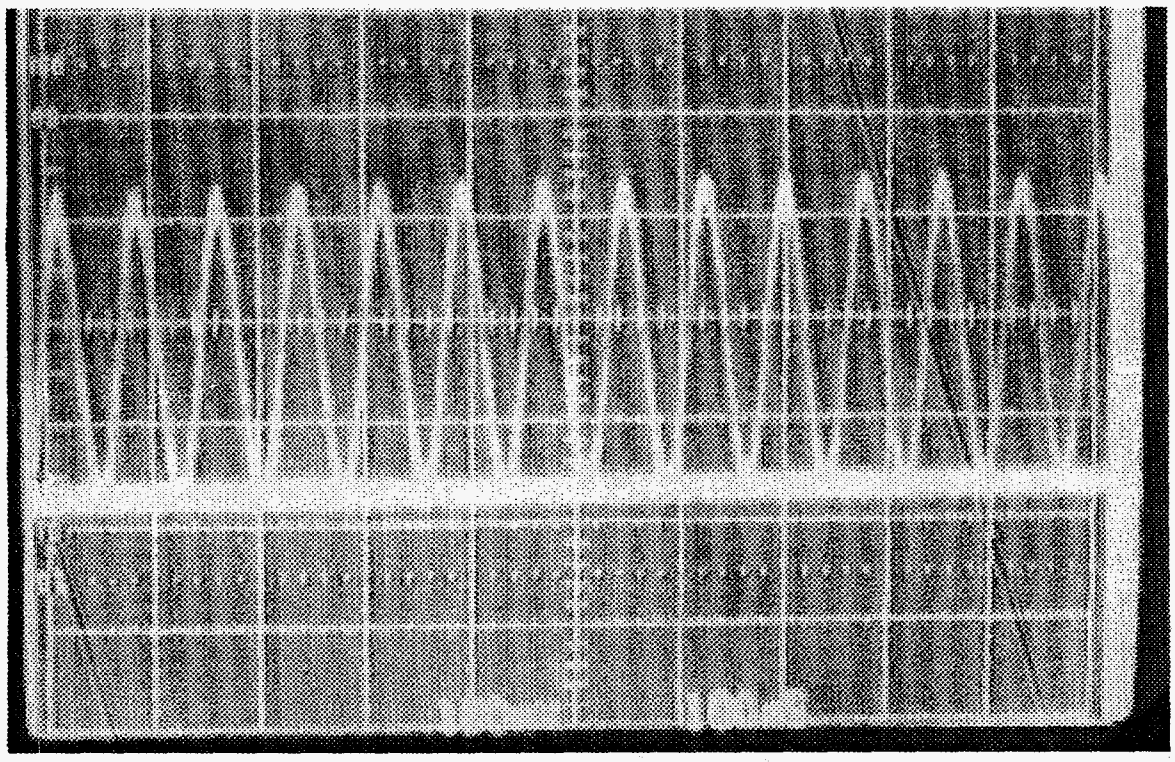

Figure 8. Correlation Output at 100 microseconds per Division.

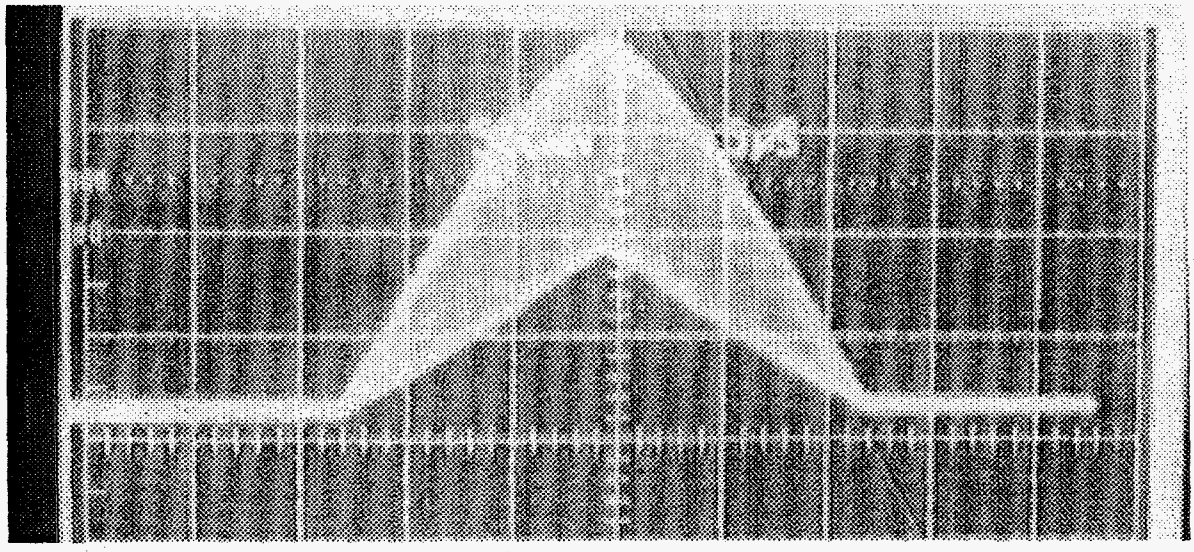

Figure 9. Correlation Output with Encoded DNA Data

Figure 9 shows the correlation output from an encoded DNA sequence which was correlated against itself. Since no inverse code was used, the maximum value corresponds to the point of highest match, i.e. when the sequences are aligned. This demonstrates the basic ability of the optical system to perform one-dimensional correlations of encoded DNA sequence data, with the results useful for sequence alignment.

The next logical step was to digitize the analog output of the detector at the proper 12.5 MHz data rate. Digitized data could be analyzed closely for finer results, e.g. misalignments between sequences on the order of a few bases, or mismatches between a few bases. To prevent aliasing effects, this would a sampling rate for the digitizer of 25 MHz to capture the entire output waveform. However, it must be remembered that the coding scheme used dictates that only every fourth correlation sample is needed, and not the entire result. As explained previously, the output of the optical correlator is already subsampled by a factor of two. Subsampling further by another factor of two results in the requisite subsampling by four. This means that the effective data rate is 6.25 MHz, thus the required sampling frequency for proper digitizing of the detector output is 12.5 MHz.

A custom analog-to-digital (A/D) circuit was built at Sandia using a high-speed A/D chip. The digital output of this circuit was then fed into a commercial image processing board, a DATACUBE MAX-SCAN board. Unfortunately, the MAX-SCAN board failed to operate properly, and a solution was not found before the termination of the project. Thus the detector output could not be digitized, and further experiments could not be performed.

Summary and Conclusion

A coding scheme was successfully developed which enables DNA sequences to be aligned using a cross-correlation operation. This was demonstrated via simulations. Attempts to implement the correlation method using an existing two-dimensional optical correlator failed due to lack of illumination power in the system. A one-dimensional optical correlator was designed and built, and was shown to perform correlations correctly. An encoded DNA sequence was successfully correlated against itself to demonstrate the basic ability of the system to perform alignment by correlation. Attempts to digitize the analog output of the correlator failed, preventing further experiments using more complicated alignment situations. Thus, the capability of the optical correlator system was demonstrated only at a basic level, and not at a level sufficient for application to real-world alignment problems.

It should be noted that during the latter part of this research, a commercial product was announced that uses a custom VLSI systolic array architecture to perform DNA sequence analysis. The advertised speed of this hardware is higher than the data rate of the optical correlator. During the course of this research, a number of significant barriers to increased data rates in the optical correlator were found, including AO cell bandwidth and linearity limitations, and the difficulty of reading data out of the system at high speed. This leads to the conclusion that the AO correlator approach should not be pursued further for DNA sequence alignment.

References

[1] K.T. Stalker, F.M. Dickey, M.L. Yee, and B.A. Kast, "Acousto-optic correlator for optical pattern recognition," in Real-time optical information processing, B. Javidi and J. Horner eds., Academic Press, Boston, 1994.

APPENDIX

LDRD Required Information

Publications/presentations resulting from the project : *none*

Invention disclosures resulting from the project : *none*

Patents resulting from the project : *none*

Software copyrights resulting from the project : *none*

Employee recruitment resulting from the project : *none*

Involvement of students in the project : *none*

DISTRIBUTION:

1	MS0188	C.E. Meyers, 4523
1	MS0843	David C. Craft, 2524
1	MS0843	K. Terry Stalker, 2524
1	MS0843	Mark L. Yee, 2524
1	MS9018	Central Technical Files, 8523-2
5	MS0899	Technical Library, 4414
2	MS0619	Review and Approval Desk, 12630 for DOE/OSTI

12 total copies