National Defence / Défense nationale

ANALYSIS OF DNA SEQUENCES BY AN OPTICAL TIME-INTEGRATING CORRELATOR: PROOF-OF-CONCEPT EXPERIMENTS

by

N. Brousseau, J.W.A. Salt, L. Gutz and M.D.B. Tucker

DEFENCE RESEARCH ESTABLISHMENT OTTAWA
TECHNICAL NOTE 92-12

Canada

May 1992
Ottawa

National Defence / Défense nationale

ANALYSIS OF DNA SEQUENCES BY AN OPTICAL TIME-INTEGRATING CORRELATOR: PROOF-OF-CONCEPT EXPERIMENTS

by

N. Brousseau, J.W.A. Salt, L. Gutz and M.D.B. Tucker
Communications Electronic Warfare Section
Electronic Warfare Division

DEFENCE RESEARCH ESTABLISHMENT OTTAWA
TECHNICAL NOTE 92-12

PCN
041LQ11

May 1992
Ottawa

ABSTRACT

The analysis of the molecular structure called DNA is of particular interest for the understanding of the basic processes governing life. Correlation techniques implemented on digital computers are currently used to perform the analysis but the present process is so slow that the mapping and sequencing of the entire human genome requires a computational breakthrough. This paper presents proof-of-concept experiments of a new method of performing the analysis of DNA sequences with an optical time-integrating correlator. Included are experimental results for the two types of analysis specified by the processing strategy. Details of the design and construction of the custom signal generators that were built to perform the experiments are presented.

RÉSUMÉ

L'analyse de la molécule d'ADN permet l'étude des fondements de la vie. Des techniques de corrélation utilisant des ordinateurs numériques sont présentement utilisées pour effectuer cette analyse mais cela est si lent que la cartographie et le séquençage de tout le génome humain exigent le dévelopement de techniques révolutionnaires. Cette note technique présente des expériences qui démontrent le concept de l'analyse des séquences d'ADN par un corrélateur optique à intégration temporelle. Les résultats expérimentaux des deux types d'analyse spécifiés par la stratégie de traitement sont présentés. La conception et la construction de générateurs des signaux spéciaux nécessaires à ces expériences sont décrites en détails.

EXECUTIVE SUMMARY

Molecular biologists tell us that each cell in our body carries all the information necessary to reconstruct the entire organism. This information is stored in a molecular structure called DNA and the analysis of DNA sequences is of particular interest for the understanding of the basic processes governing life. In that context, the mission of the Human Genome Project is to map the entire mosaic of the human DNA. In an effort to reach that objective, biochemists try to match a particular segment of DNA to existing data banks, with the possibility that the match will not be perfect. Correlation techniques implemented on digital computers are used to perform the analysis on the limited amount of data available today and the process is tedious. Considering that only a small fraction of the 3×10^9 human genome nucleotides is now available in the data banks, a mapping of the entire human genome requires a computational breakthrough.

A new method to perform the analysis of human or animal DNA sequences with an analog optical computer was recently proposed. The new method is characterized by short processing times that make the analysis of the entire human genome a tractable enterprise. The proposal is based on the utilization of a time-integrating correlator. This type of optical correlator is particularly well suited to the very fast correlation of long data streams such as the data involved in the analysis of DNA.

This technical note presents proof-of-concept experiments of the new method. Included are experimental results for the two types of analysis specified by the processing strategy. Details of the design and construction of the custom signal generators that were built to perform the experiments are presented.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont.)

LIST OF FIGURES

LIST OF FIGURES (cont)

LIST OF TABLES

LIST OF ABBREVIATIONS

DNA:	deoxyribonucleic acid
FIFO:	first in first out
IC:	integrated circuit
MIPS:	million instructions per second
PC:	personal computer
TIC:	time-integrating correlator

1.0 INTRODUCTION

The analysis of the molecular structure called DNA is of particular interest for the understanding of the basic processes governing life. All living organisms encode their genetic information in the same way, by using linear polymers of phosphoric acid and sugar (deoxyribose) upon which are attached four different bases, adenine (A), cytosine (C), guanine (G) and thymine (T).

Over the past ten years, DNA sequencing techniques have advanced sufficiently for a modest start to be made on harvesting and analyzing the formidable array of genetic diversity in life forms [1-3]. Most of the DNA sequence information available today is tabulated in the GenBank* database. Release 65 (September 1990) of this database contains 49×10^6 nucleotides from all organisms, divided into thirteen divisions. As the database grows towards its projected size of 3×10^9 for the human genome alone, it can be foreseen that current equipment will quickly become utterly impractical to use.

The problem considered in this technical note is the analysis of human or animal DNA sequences where biochemists attempt to match a query sequence of DNA to an identical or a similar segment that may be present in the existing computer databases. Correlation techniques implemented on digital computers are used to do the sequence matching on the limited amount of data available today and the process is tedious. Considering that only a small fraction of the 3×10^9 human genome nucleotides is now available and stored in the data banks, a computational breakthrough is required to allow the processing of the entire human genome.

Optical processing technique using Time-Integrating Correlators (TIC)s that could substantially reduce the analysis times have been proposed [4]. This type of optical correlator is particularly well suited to the very fast correlation of long data streams such as the data involved in the analysis of DNA. The processing times [4] for the analysis of a 50×10^6 bases database as a function of the number of bases in the query sequence are presented in Figure 1. The left of the figure uses log-log axis and covers query sequences of length 12 to 857. Semi-log axis are more convenient for the right of the figure because the analysis time varies linearly with the length of the query sequence. The abscissa and ordinate are respectively drawn on a logarithmic scale and a linear scale.

The concept of a TIC using a Mach-Zehnder architecture is illustrated in Figure 2. The beam splitter separates the incident laser beam into two paths. M1 and M2 are folding mirrors. The two beams diffracted by the Bragg cells are mixed

* Produced by GenBank c/o IntelliGenetics Inc. 700 East El Camino Real, Mountain View CA 94040.

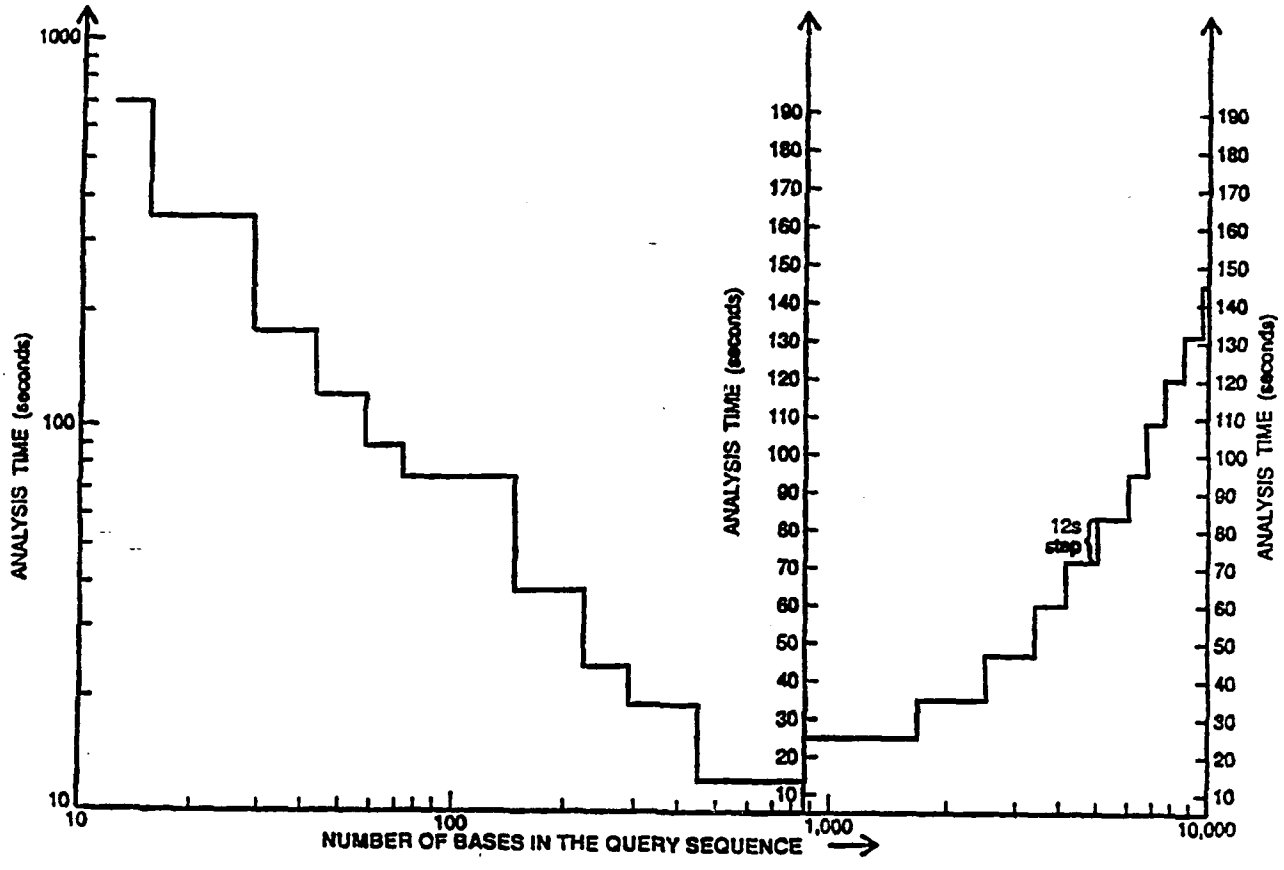


Figure 1: Processing time for the analysis of a $50x10^6$ bases database as a function of the number of bases in the query sequence.

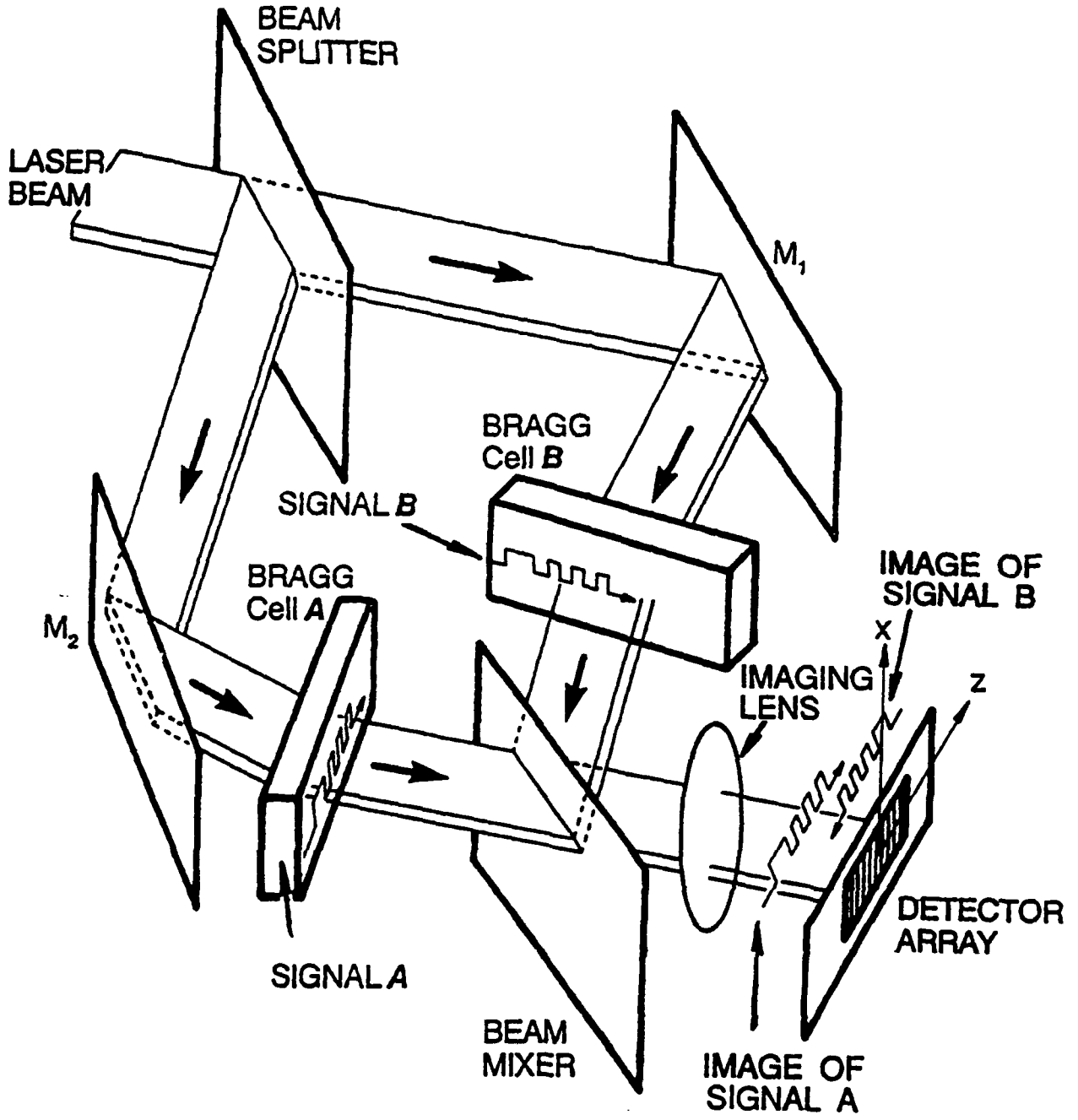


Figure 2: Time-integrating correlator: Mach-Zehnder architecture.

together by a beam mixer. The two diffracted light distributions are coaxial and imaged in such a way as to be counterpropagating on the detector array that performs a time-integration. A review of the principle of operation of TICs with emphasis on the characteristics and parameters that have an impact on the design and operation of a TIC applied towards the analysis of DNA sequences is presented elsewhere [4].

2.0 DNA ANALYSIS STRATEGY

2.1 Representation of DNA Bases

DNA sequences are built from four bases represented by the letters A, C, G and T. A fifth letter, N, is used to represent unknown elements at particular locations in a sequence. The sequences representing segments of the human genome have to be transformed into electrical signals suitable as inputs to the Bragg cells of the TIC. One way to accomplish this is to represent each base by a binary pseudorandom sequence as would be used in spread spectrum code division multiple access communications. The bits (0 and 1's) specified by the representations of the bases can be implemented using binary phase-shift-keyed modulation [5, p.16-18]. For our proof of concept experiment we use the short and long representations listed in Table 1 and 2 and in Figure 3 that have been selected for the low value of their cross-correlation. The short representations (7-bits long) were found by performing a systematic search for a set of five pseudorandom sequences having cross-correlation magnitude as low as possible. When the autocorrelation peak is normalized to 7, it is possible to find many sets of five sequences whose maximum cross-correlation value is three. However, it is impossible to find a set having a lower maximum cross-correlation value. We choose the set of five representations listed in Table 1 and illustrated in Figure 3.

Table 1: Short representations of the DNA bases where each base is represented by 7-bit pseudorandom sequences.

Base	Representation
Adenine (A)	0 0 0 0 0 0 1
Cytosine (C)	0 1 0 0 1 1 0
Guanine (G)	1 0 1 0 0 1 0
Thymine (T)	1 1 0 1 0 0 0
Unknown (N)	1 1 1 0 1 0 1

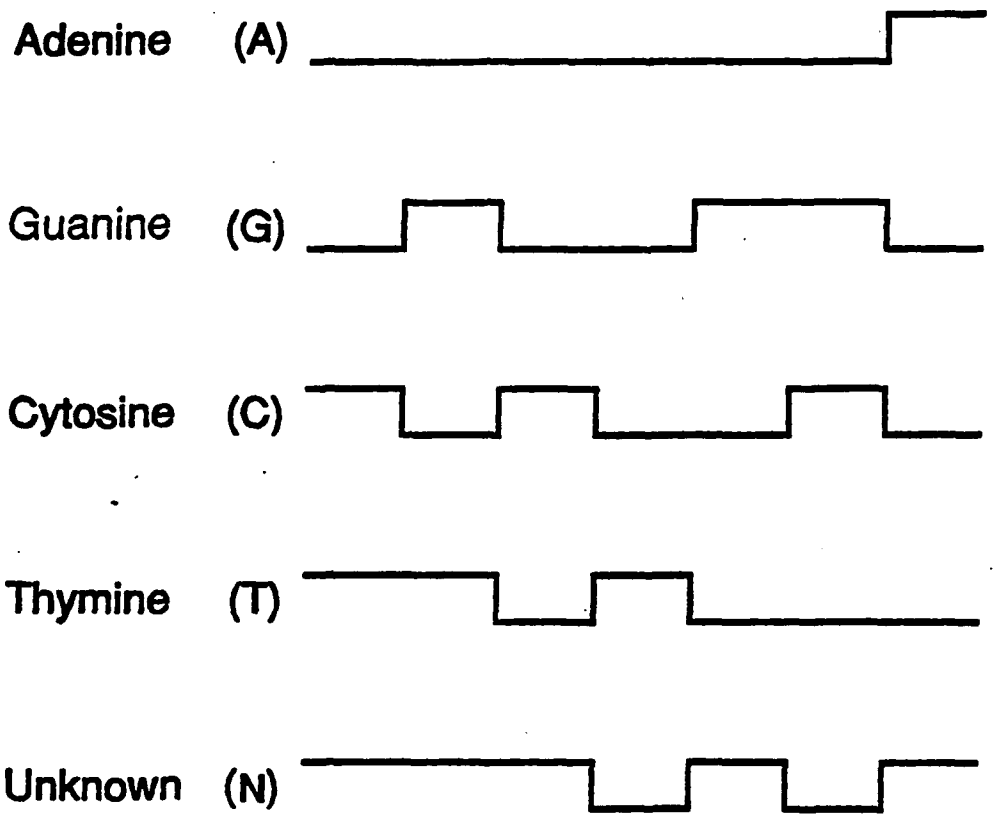


Figure 3: Short representations of the DNA bases where each base is represented by a 7-bits long pseudorandom sequence.

2.2 DNA Analysis Strategy

The purpose of this section is to briefly review the strategy to implement the analysis of a DNA sequence with a TIC that was proposed in [4]. We wish to find segments of the database that are identical or similar to the query sequence and their location within the database. We also want to produce a base-by-base comparison of the query sequence using the segments of the database that are identified as correlating with the que y sequence. The analysis is made using a two-level procedure. A coarse analysis is first used to locate the area of the database

Table 2: Long representations of the DNA bases with 255-bit maximum length pseudorandom sequences [6,p.62].

Base	Octal representation	polynomial representation
Adenine (A)	435	$x^8+x^4+x^3x^21$
Cytosine (C)	453	$x^8+x^5+x^3x+1$
Guanine (G)	455	$x^8+x^5+x^3x^21$
Thymine (T)	515	$x^8+x^6+x^3x^2+1$
Unknown (N)	537	$x^8+x^6+x^4x^3+x^2+x+1$

that are similar or identical to the query sequence. Figure 4 illustrates the coarse analysis of a DNA sequence. A database is illustrated as it propagates through Bragg cell A just before the passage of the segment that is identical to the query sequence. The signal formed by the repetitions of the query sequence is illustrated at the same moment in Bragg cell B. The correlation peak will start formation a few moments later, in about the transit time in the Bragg cell divided by two. Then, a fine analysis (see Figure 5), is performed on the database segments identified by the coarse analysis to establish a base-by-base comparison. Figure 5 illustrates the fine, base-by-base analysis of a DNA sequence. The database and the query sequence are represented by long pseudorandom sequences that almost fill the Bragg cells' apertures. The system is illustrated at the moment when the base G is correlating.

Figure 6 represents the flow of data in a DNA analysis system based on an optical TIC. On the left side the human genome data base has a potential of 3 billion bases. Currently there are approximately 50 million bases of sequence available from all living organisms. The 50 million bases that are known are stored in a digital database where they are designated by letters. These letters are then represented by pseudorandom binary sequences and transformed into analog signals which are suitable

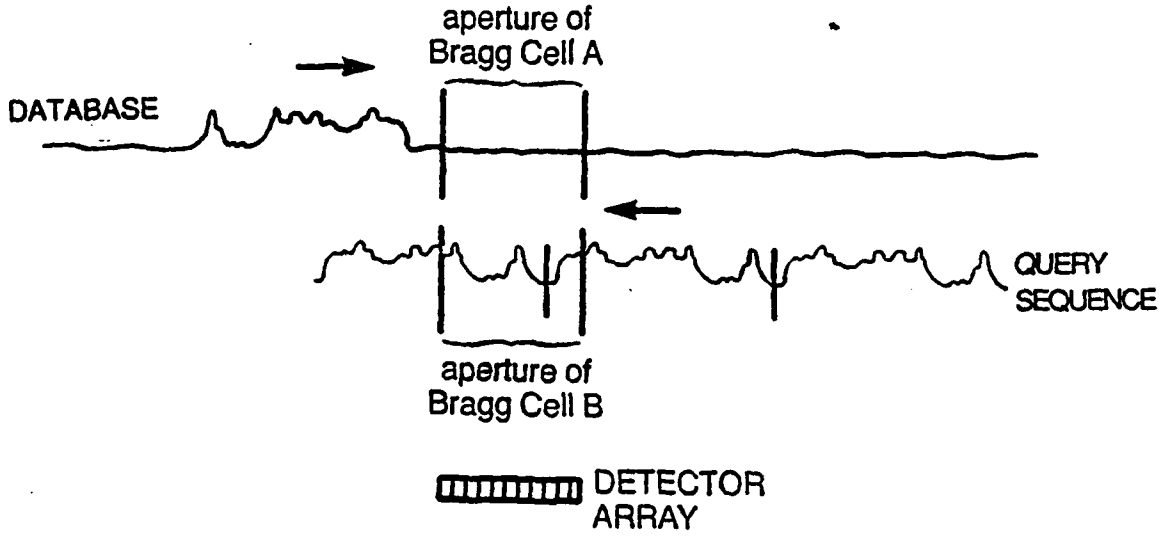


Figure 4: Coarse analysis of a DNA sequence.

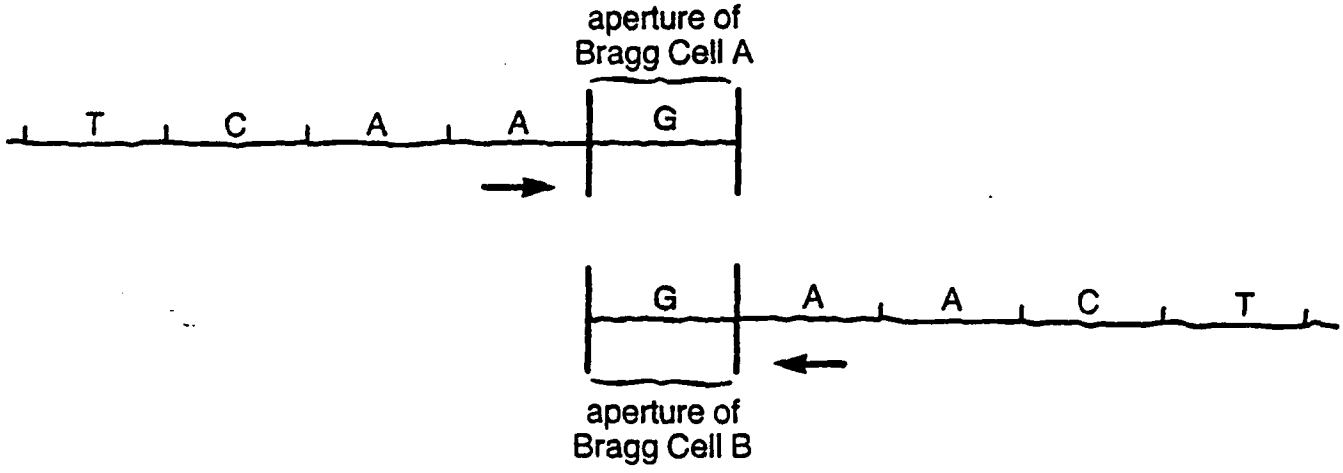


Figure 5: Fine, base-by-base analysis of a DNA sequence.

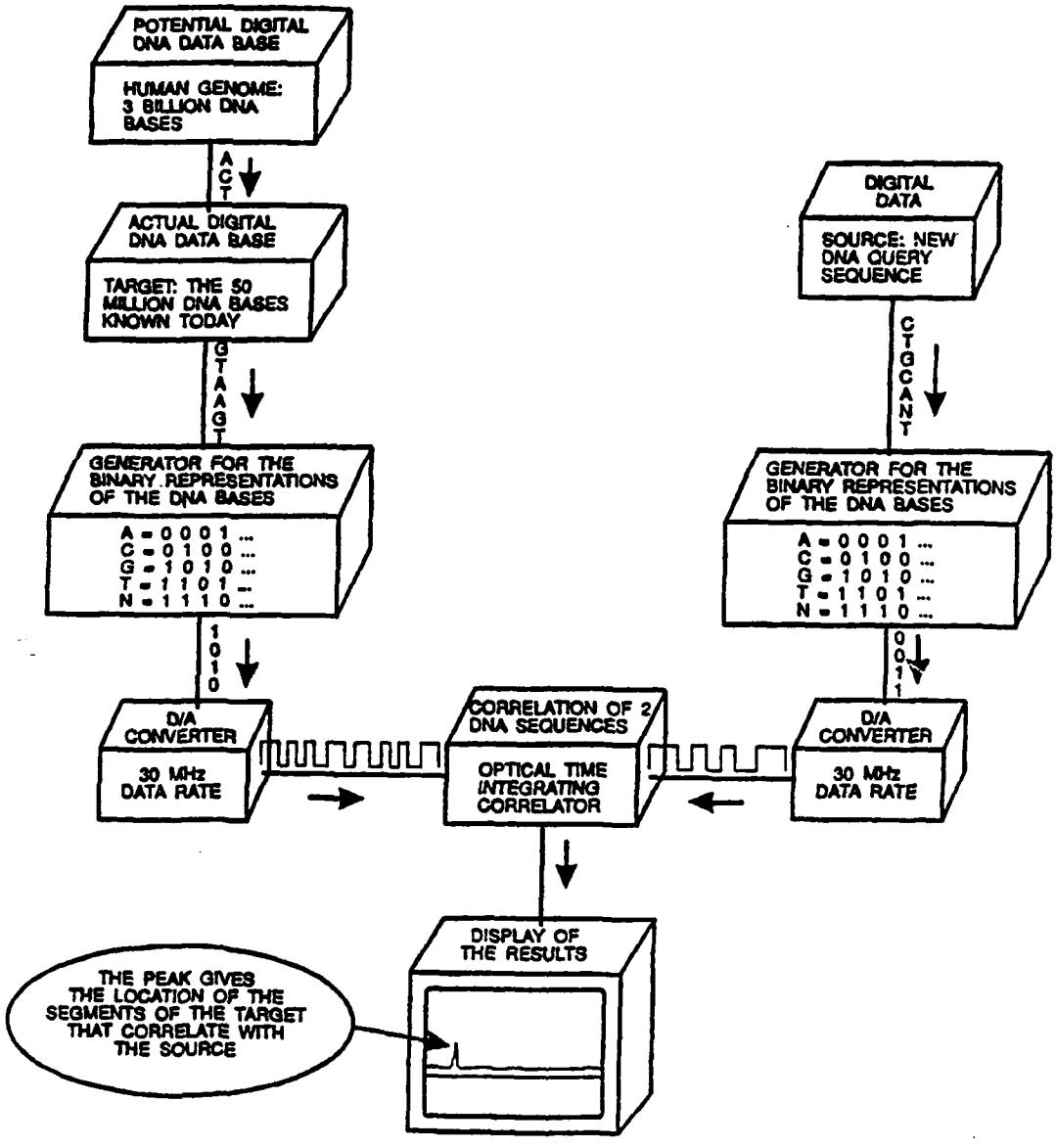


Figure 6: The flow of data in a DNA analysis system based on an optical TIC.

to operate a Bragg cell. The right side represents the new query sequence acquired by a scientist. It undergoes the same transformation and is correlated with the database from the left side by the TIC. The results are displayed and if the query sequence was not already included in the known DNA database, it is incorporated into the database.

3.0 CUSTOM DNA SEQUENCE GENERATORS

3.1 Hardware Design

In this section, a detailed description will be given of the hardware designed to carry out the proof-of-concept experiment. The custom signal generators are designed to produce two encoded DNA sequences to the TIC. The circuit boards (see Figure 7) consist of two FIFO buffers, which provide data through a parallel-to-serial converter. The larger buffer contains the database sequence that is generated once for each run. The smaller buffer contains the query sequence that is generated repetitively until the buffer is reset.

3.1.1 PC Interface

The circuit board is connected to a PC parallel port through a 25 pin connector, labelled P1 (see Figure 8). Data being sent from the PC on pins 2 to 8 of P1 (D0 to D7) is latched by U2, to insure the data will remain valid long enough to be read in by the FIFO's.

Pin 1 of P1 carries the SYNC signal which clocks data into the latch, U2. The SYNC signal controls when data is written to a FIFO IC. The LARGE (SMALL) signal which comes from pin 14 of P1 controls which FIFO is written to when the SYNC signal is asserted. The RESET signal goes to the RESET or CLEAR pins of most of the IC's on the board.

Pin 10 of P1 is the acknowledge pin. The signal from pin ten is sent back to the PC to indicate that the data has been received.

3.1.2 The Large FIFO

The large FIFO actually consists of four IDT 7M206 IC's (U6, U7, U8, and U9), daisy chained to form a 64k x 9 bits FIFO unit (see Figure 9). Data is written to the large FIFO when the LARGE signal is asserted and the SYNC signal goes high.

Data is read from the large FIFO when the READ signal transitions from high to low. Since U6, U7, U8, and U9 act as a single FIFO chip, the first data word written to the large FIFO is the first word to be read. Each data word is made up of 8 bits (the D8 and Q8 pins on each FIFO chip are not connected).

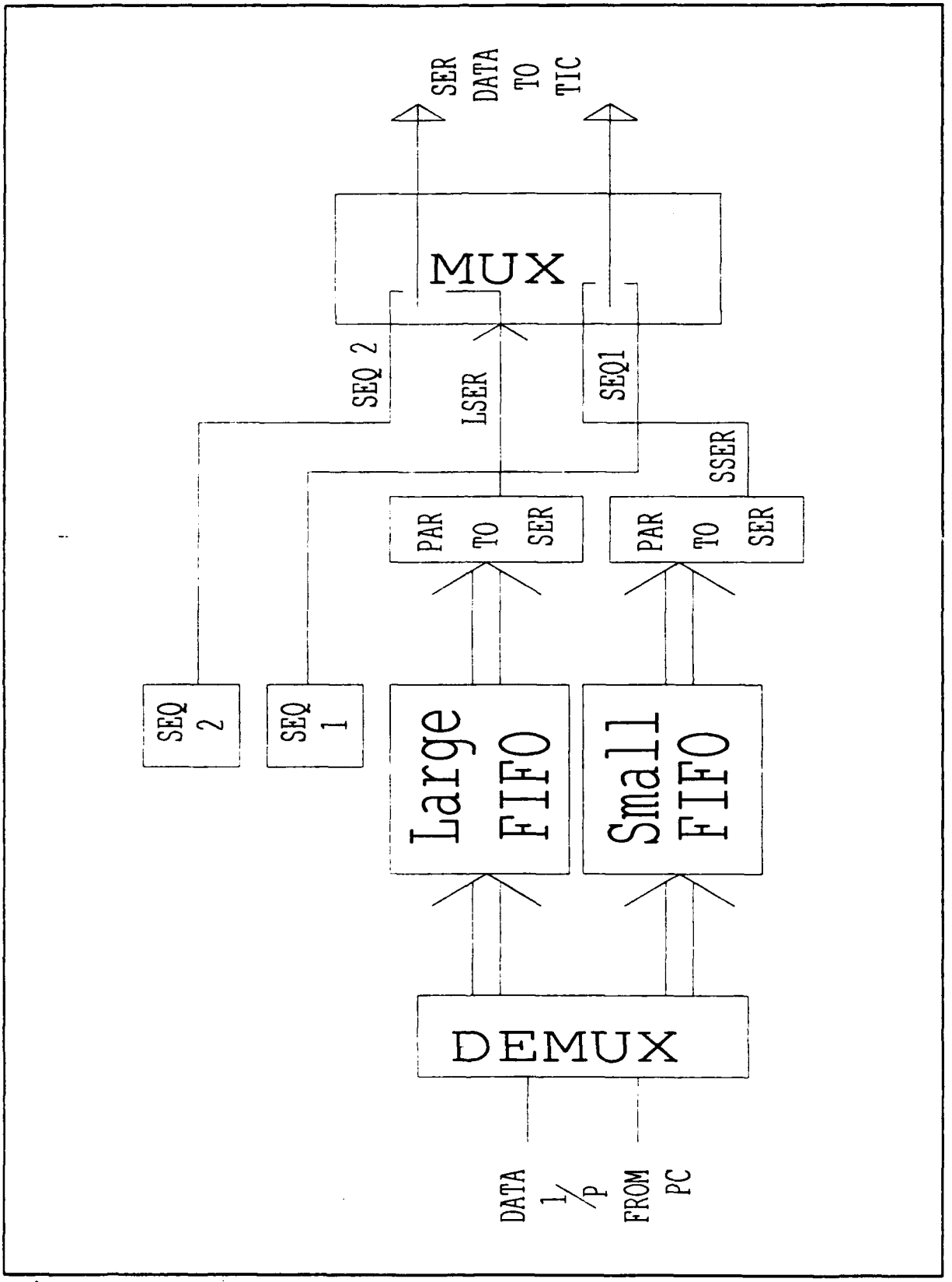


Figure 7: Circuit boards for the custom signal generators

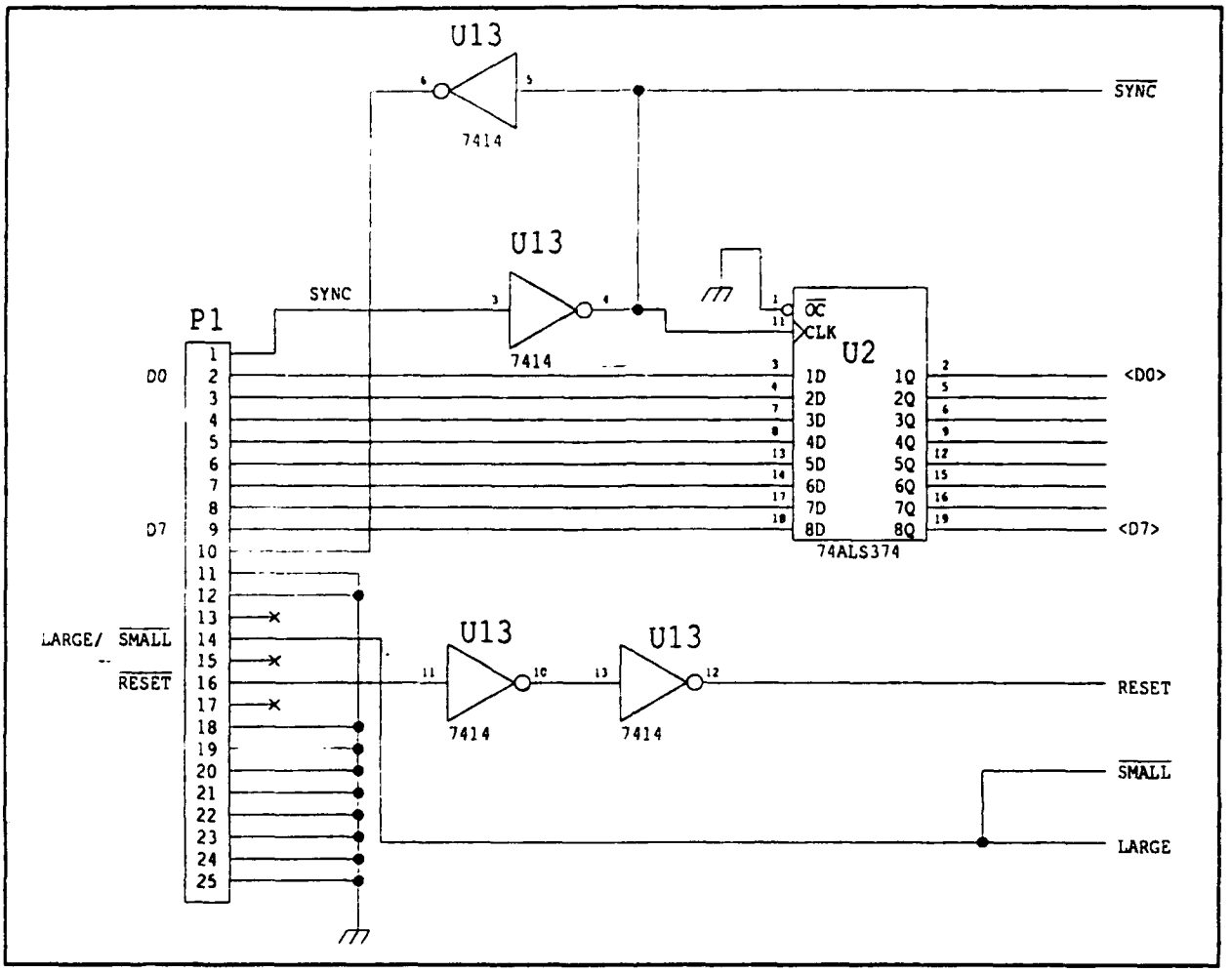


Figure 8: Connection of the circuit board to a PC parallel port through a 25-pin connector.

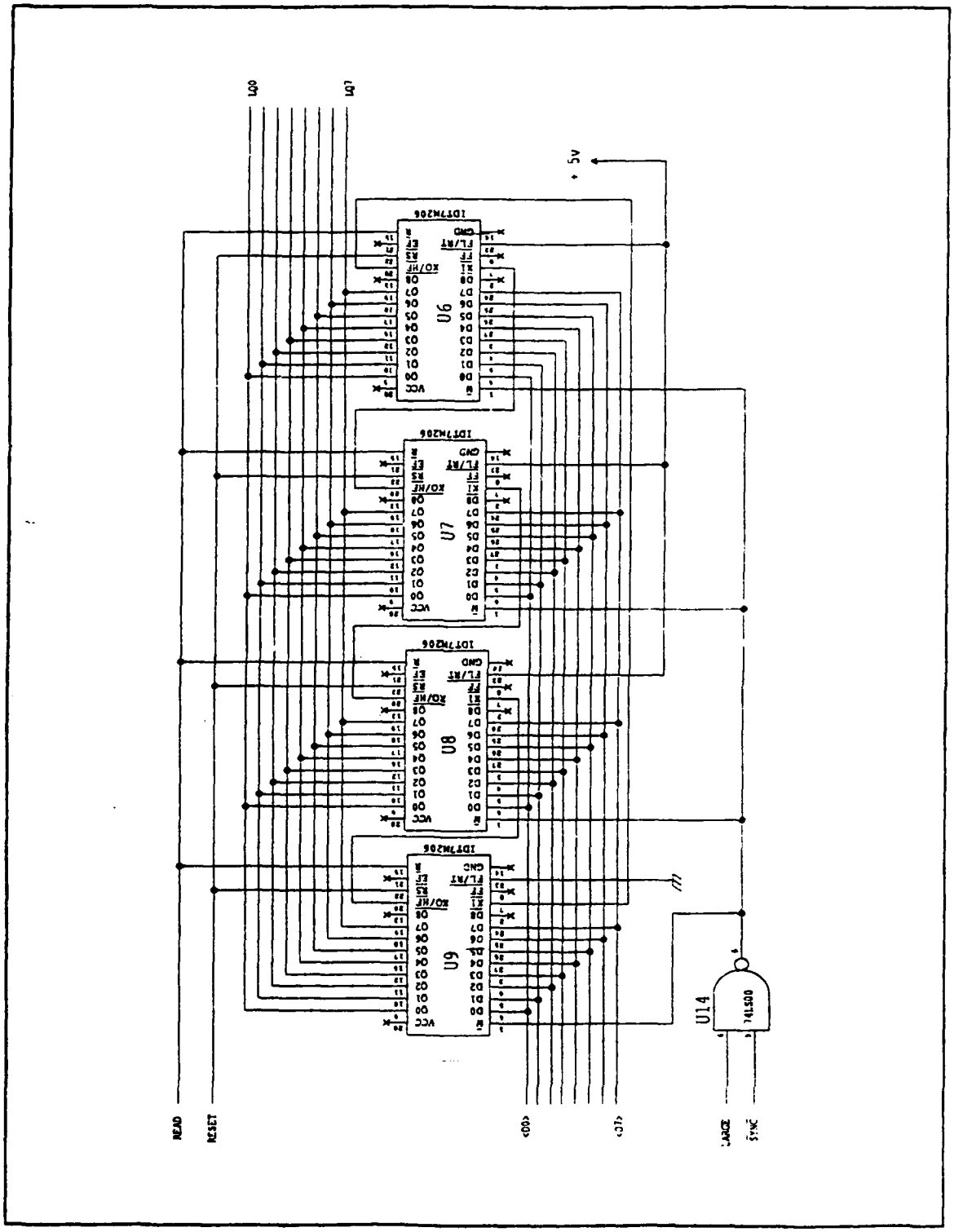


Figure 9: The large FIFO consists of four IDT 7M206 IC's (U6, U7, U8 and U9), daisy chained to form a 64k x 9 bits FIFO unit.

The large FIFO has two independent internal pointers, a read pointer and a write pointer. All the logic involved in implementing a FIFO queue is internal.

3.1.3 The Small FIFO

The small FIFO is a single IDT 7M206 IC. It is a 16k x 9 bit FIFO queue with internal read and write pointers. It also has a retransmit flag which is used to cause the data in the small FIFO to be transmitted over and over again in cyclic fashion.

Sending a low level to the retransmit (FL/RT) pin causes the read pointer to be reset to the first word which was written to the small FIFO. The retransmit pin is driven low by the Empty Flag (EF), which indicates that all the data has been transmitted. The signal is held low with a monostable multivibrator, U15 (see Figure 10) with a designed period of 30ns.

Data is written to the small FIFO when the SMALL signal is low and the STROBE signal goes high. Data is read from the small FIFO whenever the READ signal goes from high to low.

3.1.4 The System Clock

The circuit board uses an external source to generate a clock signal. Some IC's are connected directly to the external clock (EXCLK). Others require that the clock be enabled after a signal from the TIC. This signal is called INCLK. There is also a divider (U10) which divides the frequency of the clock signal by seven (switch open) or eight (switch closed), depending on the setting of switch S1. The output of the divider U10 is the READ signal (see Figure 11).

3.1.5 The Pseudo-Orthogonal Sequence Generators

After the RESET signal has been driven low by the PC, but before the BEGIN signal has been sent, the circuit board sends a sequence out through each output channel. The two sequences are chosen such that they have a low cross correlation magnitude and therefore would not interfere with the valid data sent from the two FIFO's after the BEGIN signal is asserted.

That precaution is necessary because the output light distribution produced when pseudorandom signals are present in the Bragg cells is not the same as when other types of signals are used. When the BEGIN signal is sent, it takes 50 μs for the pseudorandom signal representing the DNA sequences to fill up the Bragg cells. During that 50 μs transition time, bad data accumulates on the detector array. If the integration time is 200 μs, the first 50 μs contributes bad data and the last 150 μs contributes good data to the first frame. The second frame contains only good data, so the subtraction of these two frames, that is designed to remove the pedestal, produces a high level of residual signal because of the difference between the first and

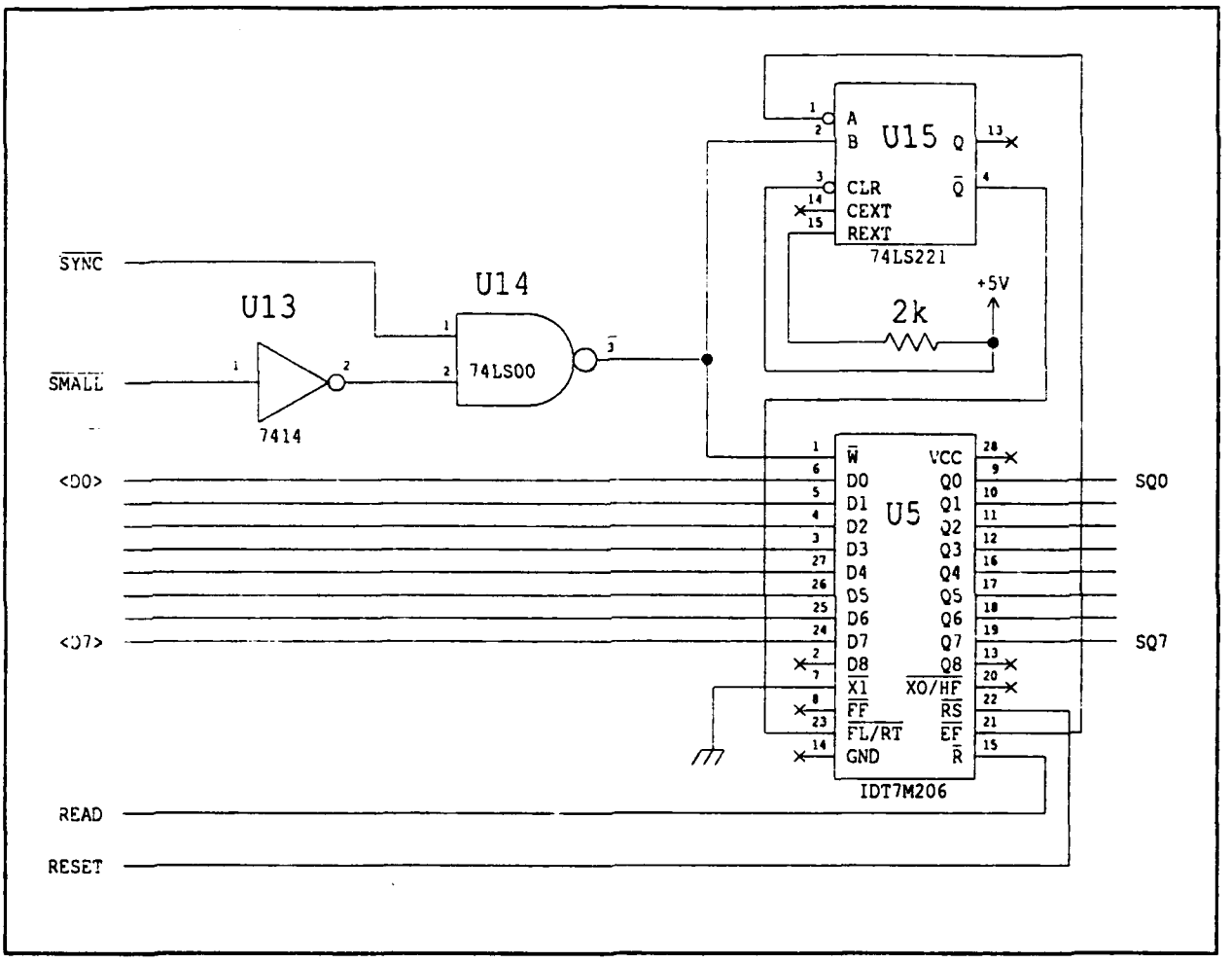


Figure 10: The small FIFO is a single IDT 7M206 IC. It is a 16k x 9 bit FIFO queue with internal read and write pointers.

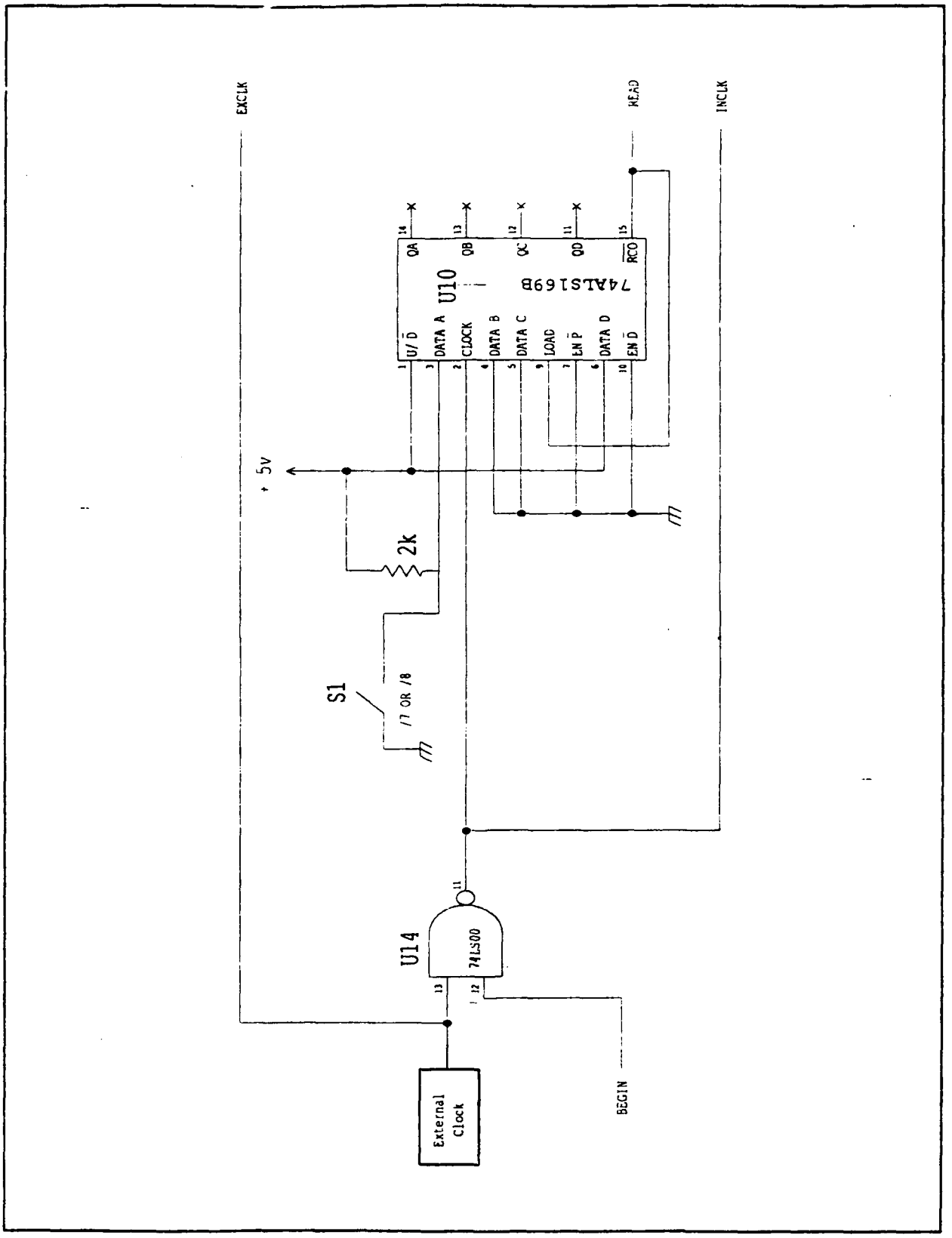


Figure 11: The system clock for the circuit board.

the second frame. That residual signal is higher than the peak detection threshold, so a peak is detected and the system, that is designed to stop after the detection of the first peak, stops there, making it impossible to perform the DNA experiment. As it is not practical to change the behaviour of the system, this problem is circumvented by having low cross correlation pseudorandom signals of the same nature than the DNA sequences circulating in the two Bragg cells. This allows proper pedestal subtraction to be performed on the first two frames of data collected by the TIC.

The first sequence (SEQ1) is a simple clock signal at half the frequency of the external clock. It is implemented with a single D-type flip-flop which is wired so that the output pins change state from high to low after each positive edge sent to the clock pin (see Figure 12a).

The second sequence (SEQ2) is also a clock signal, but it has a frequency of one quarter that of the external clock. This signal is generated using two D-type flip-flops (see Figure 12b).

The first sequence can be represented, in binary numbers, as 10101010... The second sequence would then be 11001100... These two sequences produce very little correlation between them.

3.1.6 Parallel to Serial Converters

There are two parallel to serial converters on the board (U11 and U12), one which accepts data from the large FIFO (Figure 13a) and one which accepts data from the small FIFO (Figure 13b).

Eight bits of data are read from a FIFO into each converter whenever the READ signal is high. Each time the INCLK signal is sent, one bit of data is sent out through the serial out pin. If the divider (U10, see Figure 11) is set to divide the frequency by eight, then all eight bits of data which are read from the FIFO are sent out serially. However, if the divider is set to divide by seven, the data sent to D0 will be lost and never sent out serially. In this case only D1 to D7 are sent out.

The purpose of the removal of bit D0 is to make it possible to use seven bit words to represent the DNA bases, allowing a faster system but at the expense of a small deterioration in the correlation function.

3.1.7 Output Control Logic

After the RESET signal has been sent, the board sends data out through channels A and B from the two sequence generators (SEQ1 and SEQ2) until the EXBEGIN signal is received from the TIC. The EXBEGIN signal is a pulse which activates a D-type flip-flop, causing the Q output to go from low to high (see Figure 14).

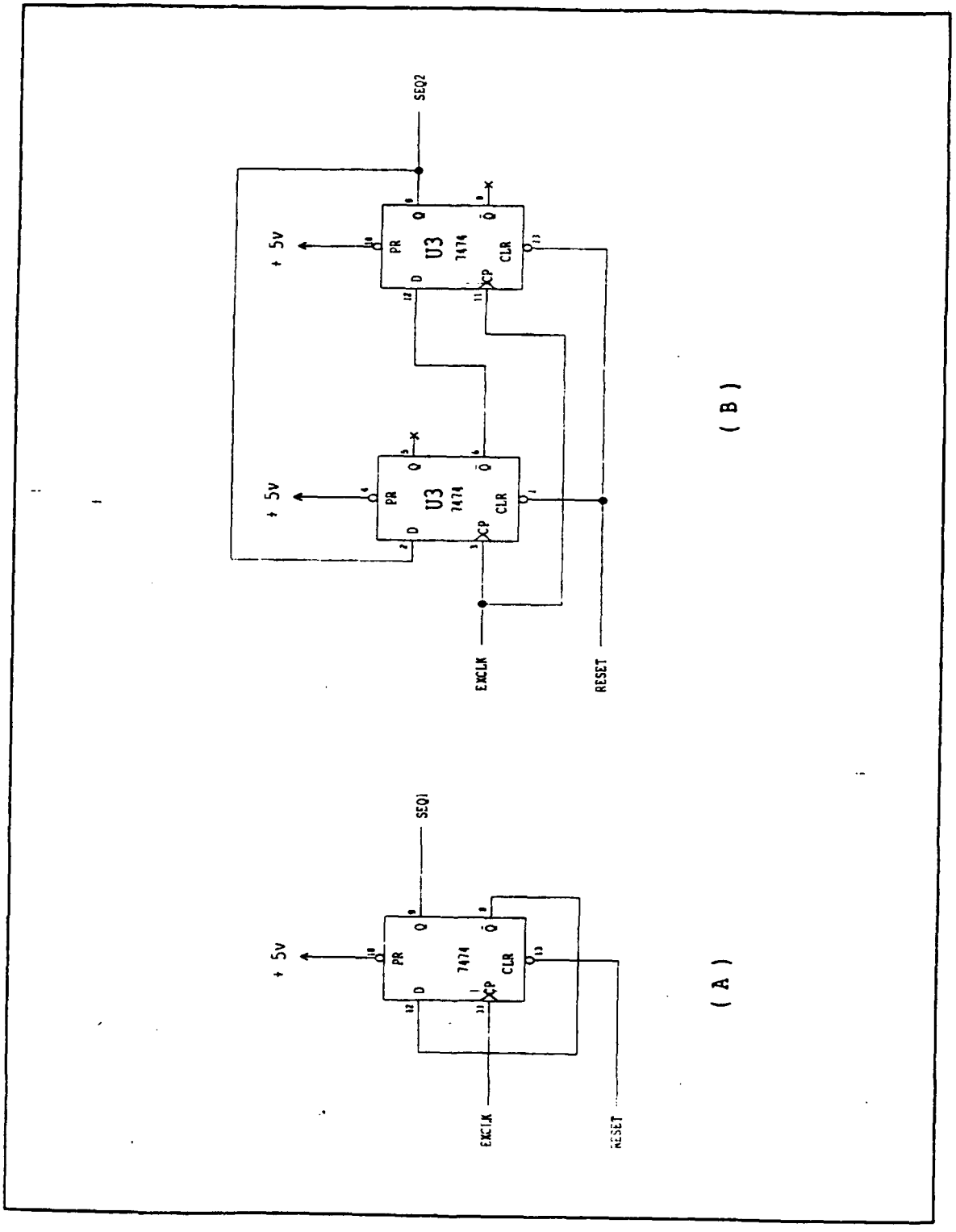


Figure 12: The two pseudo-orthogonal sequence generators. The two sequences exhibit very low cross-correlation.

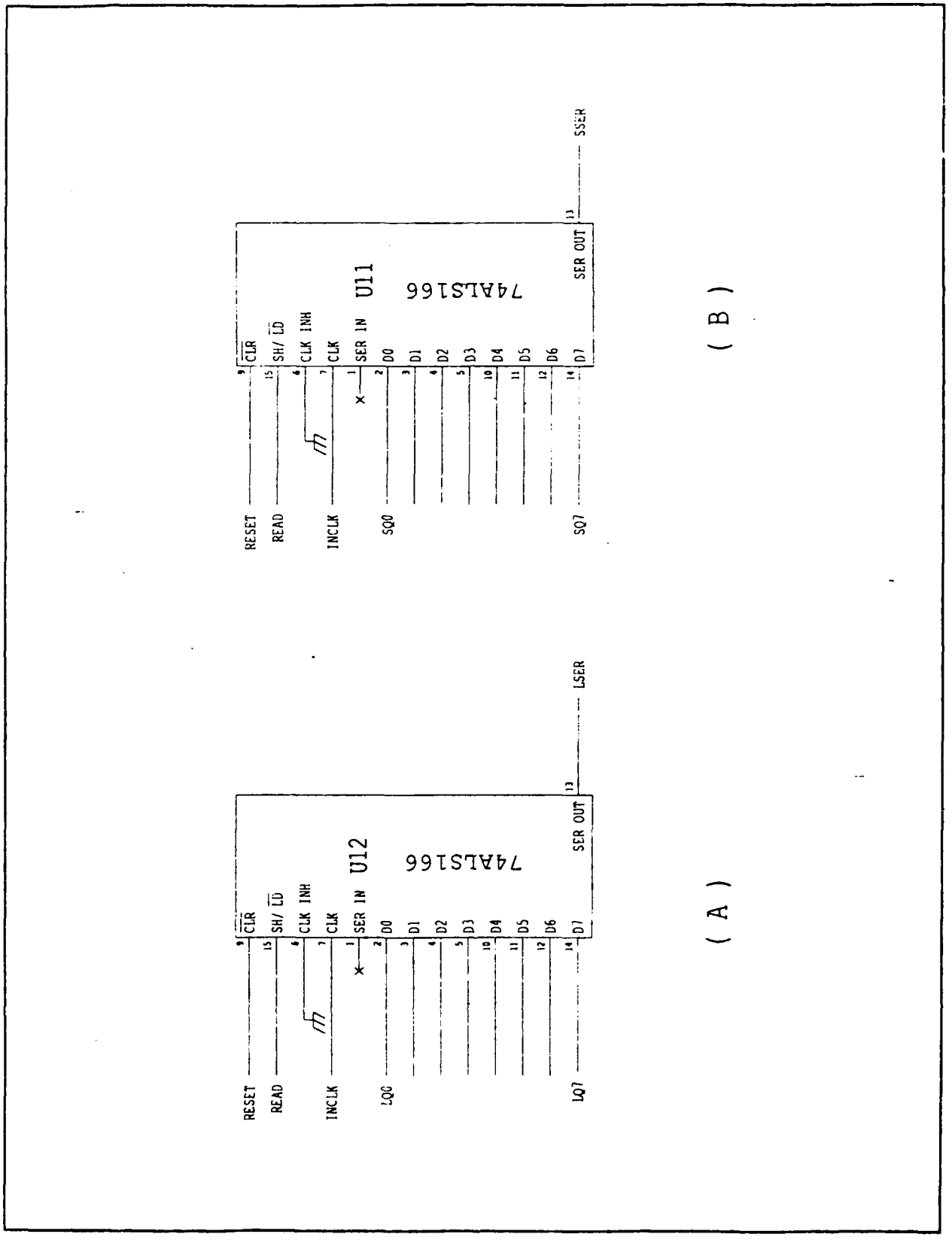


Figure 13: The two parallel to serial converters for the data from the large and the small FIFOs.

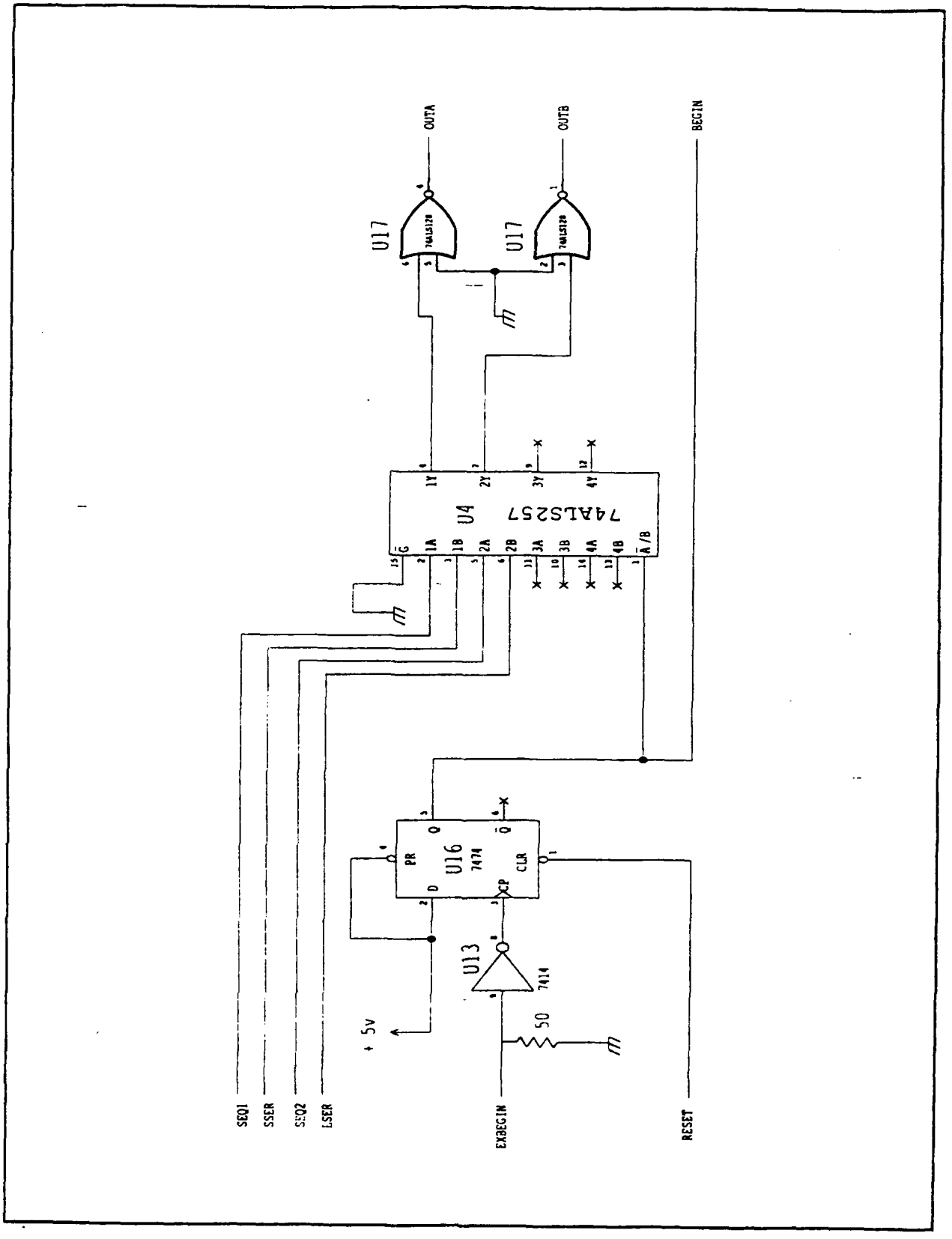


Figure 14: Output control logic for the data sent to the two Bragg cells.

The BEGIN signal is taken from the Q output of the flip-flop. This same signal also instructs the switching IC (U4) to stop loading data from the sequence generators and start sending serial data from the large and small FIFO's. All data passes through a 50 ohm line driver before being sent out one of the two output channels of the circuit board.

3.2 Software Design

The custom software to control the buffers is designed to send two data files to them and wait for further instructions. A signal from the Controller instructs the buffers when to begin outputting the sequences. The custom software also controls the function which provides a time-shift of the query sequence by increments of 21 μs. This allows displacement of the correlation peak in order to obtain a peak located within the TIC time-delay window. This function is useful for query sequences that are longer than the 80 μs time-delay window of the TIC.

3.2.1 Converting DNA Files

DNA files are text files made up of any combination of five letters: A, C, G, T, and N. These letters represent the bases in a chain of DNA. The ASCII representation of these characters are not ideal for use with the TIC, so each character must be converted into a binary vector.

Because of the design of the circuit board, the vectors must be of length seven or eight, and consist entirely of 0's and 1's. If the vectors are of length seven, switch S1 on the board must be in the open position, otherwise S1 must be in the closed position (see sections 2.4 and 2.6).

The five vectors chosen were selected such that they would produce a correlation peak only when two vectors which represent the same base are correlated. Care was taken when choosing the vectors to insure that combinations of vectors would not correlate with any other combination of vectors which did not represent the same sequence of bases of DNA (see Table 1).

3.2.2 Controlling the Circuit Board

The control signals which the CORR software is capable of sending can reset the board, write data to a FIFO, and control which FIFO is being written to. All signals to the circuit board must be sent through the parallel port of the PC.

The circuit board is reset when the program is started. The two data files are then sent to the two FIFO's. Upon completion, the program waits for further instructions from the user.

3.2.3 Sending Data to the Circuit Board

When the program is first run, it asks for the frequency at which the external clock is oscillating in MHz. The first time the data is sent to the board, it is sent exactly as it appears on disk. The program then asks the user to either quit the program, send the same data which was just sent, or shift the data.

If the data is to be shifted, the long file is sent to the large FIFO in the same way it was originally sent. However, the data being sent to the short file is altered. The first character sent to the small FIFO is not the first character in the short file. Instead the program skips over the number of vectors corresponding to the specified shift, taking into account the speed of the external clock. If a second shift is then performed, twice as many vectors will be skipped, and so on. Vectors that are skipped are appended to the end of the file.

The purpose of the shifting feature is to change the relative time delay of the two input signals. Different time delays, corresponding to slightly overlapping windows of the correlator, can then be tried until the proper time delay is found in which the correlation peak is visible. This allows the correlation peak to be brought inside the time-delay window of the correlator.

4.0 DEMONSTRATION OF THE COARSE ANALYSIS

4.1 Description of the Coarse Analysis

The purpose of the coarse analysis is to find the areas of the database that are similar to the query sequence. The process involved in the production of the correlation peaks for the coarse analysis consists of sending the database sequence continuously through Bragg cell A (see Figure 2). Simultaneously, the query sequence is passed through Bragg cell B. The output of the detector array is examined at regular intervals T. The pedestal is removed and the presence of a peak is verified by comparison with a preset threshold level for each collected frame. The setting of the threshold level determines the degree of similarity that is required to declare that a certain segment of the database correlates with the query sequence. The higher the peak, the better the correlation between the query sequence and the database. These operations can be performed in real time with proper hardware implementation. When a segment of the database in Bragg cell A is identical or sufficiently similar to the query sequence in Bragg cell B, correlation peaks will be produced and detected. The time of occurrence of such events is associated with the position of the query sequence in the database and can be determined by knowing which frame contains the correlation. All

of the occurrences of a correlation peak will be noted and the fine analysis will follow to obtain a base-by-base comparison of the query sequence with the database.

4.2 Proof of Concept Experiment for the Coarse Analysis

4.2.1 Introduction

The feasibility of performing fast DNA analysis with a TIC was experimentally demonstrated. The experimental system included a TIC and a Controller originally designed to process spread-spectrum communication signals. In our experiment, the spread-spectrum signal generators were replaced by the custom-built signal generators producing bit streams representing DNA sequences that are described in Section 3.0. The Controller could not be modified to accommodate the special requirements of DNA analysis so we had to design our experiments to circumvent a few bothersome idiosyncrasies. For example, during analysis the system was designed to stop after the detection of a peak; this caused loss of synchronization, prevented contiguous operation and made it impossible to verify the presence of the series of correlations normally produced by long query sequences. Consequently, we had to limit our experiments to the detection of the first peak.

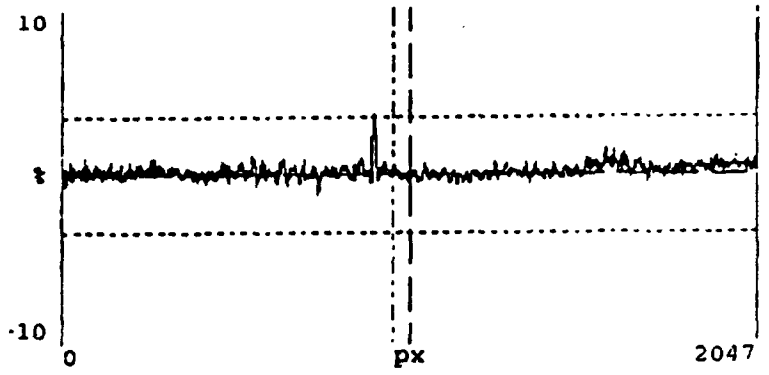
However, it was sometimes possible to detect a second peak of a correlation pattern by executing two consecutive runs with increasing threshold levels. This occurred when the start of the correlation between the query sequence and the corresponding segment of the database was not synchronized with the beginning of a collection frame of the detector array. As a result a small correlation peak (see Figure 15a) was produced in the frame during which the correlation did not exist for the entire collection time. The subsequent frame produced a maximum height correlation peak (see Figure 15b).

4.2.2 Parameters of the Experimental System

The TIC used to perform the experiment included two Bragg cells made of TeO_2 with 40 μs effective apertures. The time-delay window of the TIC was consequently 80 μs. The system was operated at a 3 MHz data rate, and each base was represented by a 7-bit pseudorandom sequence (see Table 1 and Figure 3). The phase-shift pedestal removal technique was used and the effective integration time T for a frame of data was 416 μs, the shortest available with this system.

4.2.3 Parameters of the Query Sequences and of the Databases

The bases of the database were produced by a random generator in the custom software. The query sequence was produced by selecting a known 270-bases segment from the database content. The database sequences contained 20460 bases with representations that were 7-bits in length and performed the analysis at a bit rate of 3 MHz, giving a database duration of

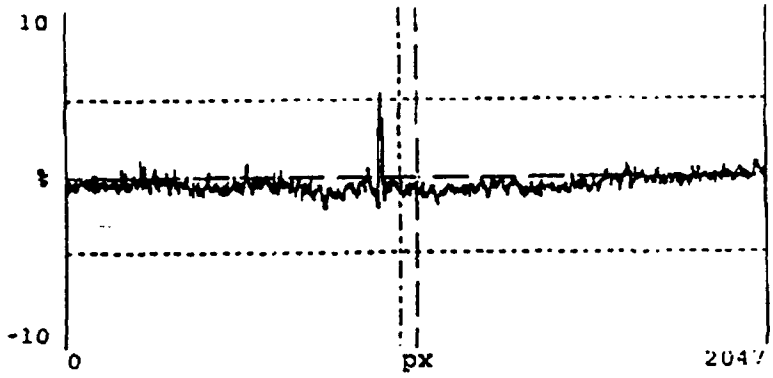


(a)

source location:
between bases 270
and 540

peak frame number: 2

peak detection
threshold: 3.5%

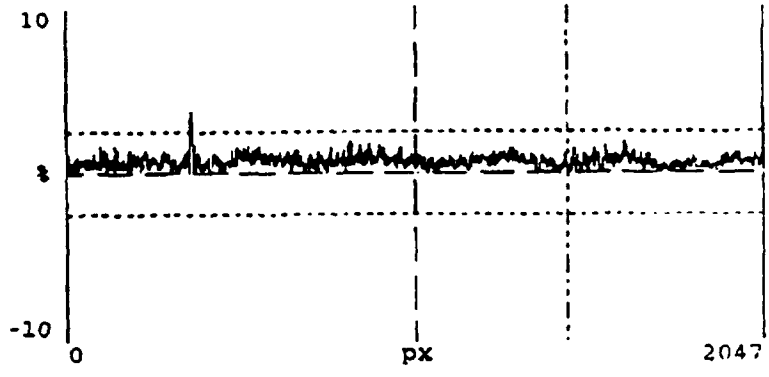


(b)

source location:
between bases 270
and 540

peak frame number: 3

peak detection
threshold: 5.5%



(c)

source location:
between bases 2970
and 3240

peak frame number: 18

peak detection
threshold: 2.5%

Figure 15: Correlation peaks produced by a query sequence located between bases 270 and 540 in the database.

47.8 ms. The query sequences contained 270 bases which, at 3 MHz, had a duration of 630 μs. The time-delay window of the TIC displayed an 80 μs section of the correlation function at a time. Thirty time-delay increments, each 21 μs long, were necessary to go through the entire query sequence.

4.2.4 Location of a Query Sequence in the Database

A segment of the database was selected as a query sequence and its location in the database was determined. The number of frames that had to be processed before detecting a peak was calculated. Having this information made it possible to check that the peak was detected at the right frame number. Query sequences from different locations in the database were tried and the peaks appeared in the expected frames. Figures 15a and 15b illustrate a correlation peak produced by a query sequence located between bases 270 and 540 in the database. In this case the setting of different thresholds levels for two runs permitted observation of the first two peaks. Figure 15c shows a correlation peak produced by a query sequence located between bases 2970 and 3240 of the database.

4.2.5 Correlation of a Query Sequence Similar to a Segment of the Database

The second experiment consisted of detecting query sequences that were only similar to a particular segment of the database. Figure 16a illustrates a correlation peak produced by a query sequence identical to a segment of the database located between bases 360 and 720. Figures 16b, 16c and 16d illustrate the peak where 20%, 40 % and 50% of the bases of the query sequence have been changed. The amplitude of the correlation peaks matches the number of bases that are common to the query sequence and the database.

In the above experiments each run was executed in 47.8 ms, the time it took for the 20460-bases database to go through a Bragg cell.

5.0 DEMONSTRATION OF THE FINE ANALYSIS

5.1 Description of the Fine Analysis

The purpose of the fine analysis is to produce a base-by-base comparison between the database and the query sequence. The presence of any discrepancies will be revealed with all the details of these features. The key to the fine analysis is to use lower data rates, representations of the bases that are much longer and to perform the analysis only on the segments of interest identified by the coarse analysis. Maximum-length sequences containing 255 bits and an integration time of 255 μs were used with a data rate of 1 MHz. When the TIC operates in this mode (see Figure 5), the correlation of the bases of the database should be synchronized with the bases of the query sequence to optimise the height of the correlation peaks. The controller of the system and the access to the memory

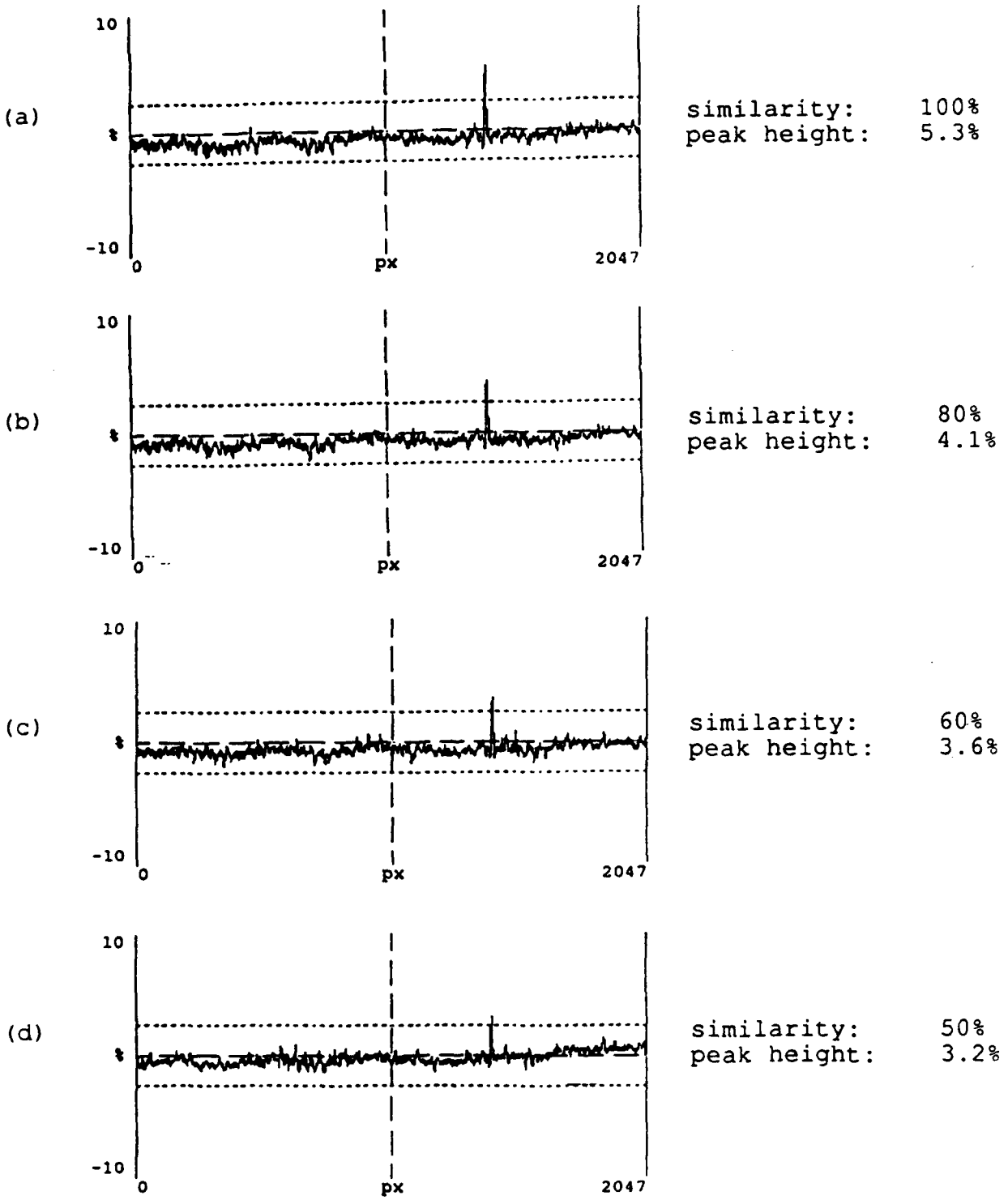


Figure 16: Correlation peaks produced with query sequences having a certain degree of similarity with the database. a-100% similarity ; b-80% similarity; c-60% similarity and d-50% similarity.

containing the query sequence and the database should be designed with enough flexibility to provide the capability to move back and forth in the memory in order to analyse in detail the gaps and discrepancies between the query sequences and the database.

The proof of concept experiments were performed to demonstrate the capabilities of the TIC to recognize individual bases. It is the building block of the fine analysis whose purpose is to establish a base-by-base comparison of the query sequence with the database. These experiments were performed with the TIC, the Controller and the two spread-spectrum signals generators that are usually integrated to our system for the processing of communication signals. Pseudorandom maximum-length sequences of length 255 bits were used to represent the bases (see Table 2) and an effective integration time of 510 μs was used to integrate twice over the sequences using the phase-shift pedestal removal technique. Twenty randomly selected bases were used as a database sequence and sent by one of the signals generators to Bragg cell A when requested by the operator. Each base was labelled t1 to t20, as indicated in Table 3 and Table 4, to facilitate the discussion. A 7-bases long segment from the database located between bases t4 and t10 was selected as a query sequence. The bases in the query sequence were labelled s1 to s7. The length of the database and of the query sequence were chosen to be very short because these experiments were not automated.

5.2 Fine Analysis of a Query Sequence Identical to a Segment of the Database

The presence in a 20-bases long database of each of the bases of a 7-bases long query sequence was confirmed. The first base of the query sequence, s1, was successively correlated to the first three bases of the database t1, t2 and t3 (see Table 3) and no correlation peak was obtained. When s1 was correlated with the fourth base of the database t4, a correlation peak was obtained thus defining the beginning of the segment of the database containing the query sequence. The following bases of the query sequence, t5, t6, t7, t8, t9 and t10 were respectively correlated with the bases s2, s3, s4, s5, s6 and s7 of the database and correlation peaks were obtained each time. Figure 17 shows the ten correlation patterns obtained during that experiment. It was then concluded that the query sequence was identical to the segment t4-t10 of the database.

5.3 Fine Analysis of a Query Sequence Similar to a Segment of the Database

The bases located at position t6 and t9 of the database were changed from C and N to T and A and the same procedure was followed. The first base of the query sequence, s1, was successively correlated to the first three bases of the database t1, t2 and t3 (see Table 4) and no correlation peak was obtained. When s1 was correlated with the fourth base of the database t4, a correlation peak was obtained thus defining the beginning of the

Database: t1 t2 t3 t4 t5 t6 t7 t8 t9 t10 t11 t12 t13 t14 t15 t16 t17 t18 t19 t20
T G A C N C A N N G T G A N N A G G G T

Query sequence: s1 s2 s3 s4 s5 s6 s7
C N C A N N G

position of the bases in the database	position of the bases in the query sequence	
	1st phase of the analysis / 2nd phase of the analysis	results of the correlation: are the bases identical?
t1 T	s1 C	NO
t2 G	s1 C	NO
t3 A	s1 C	NO
t4 C	s1 C	YES
t5 N	s2 N	YES
t6 C	s3 C	YES
t7 A	s4 A	YES
t8 N	s5 N	YES
t9 N	s6 N	YES
t10 G	s7 G	YES
t11 T		
t12 G		
t13 A		
t14 N		
t15 N		
t15 A		
t17 G		
t18 G		
t19 G		
t20 T		

Table 3: Correlations produced by a fine analysis performed with a 7-bases query sequence contained in a database that is 20-bases long in which a segment is identical to the query sequence. The region where a match is found is between position 4 and 10 of the database.

Database: t1 t2 t3 t4 t5 t6 t7 t8 t9 t10 t11 t12 t13 t14 t15 t16 t17 t18 t19 t20

T G A C N T A N A G T G A N N A G G G T

Query sequence: s1 s2 s3 s4 s5 s6 s7

C N C A N N G

position of the bases in the database	position of the bases in the query sequence	
	1st phase of the analysis / 2nd phase of the analysis	results of the correlation: are the bases identical?
t1 T	s1 C	NO
t2 G	s1 C	NO
t3 A	s1 C	NO
t4 C	s1 C	YES
t5 N	s2 N	YES
t6 T	s3 C	NO
t7 A	s4 A	YES
t8 N	s5 N	YES
t9 A	s6 N	NO
t10 G	s7 G	YES
t11 T		
t12 G		
t13 A		
t14 N		
t15 N		
t15 A		
t17 G		
t18 G		
t19 G		
t20 T		

Table 4: Correlations produced by a fine analysis performed with a 7-bases query sequence contained in a database that is 20-bases long in which a segment is similar to the query sequence. The region where a match is found is between position 4 and position 10 of the database with discrepancies at location 6 and 9.

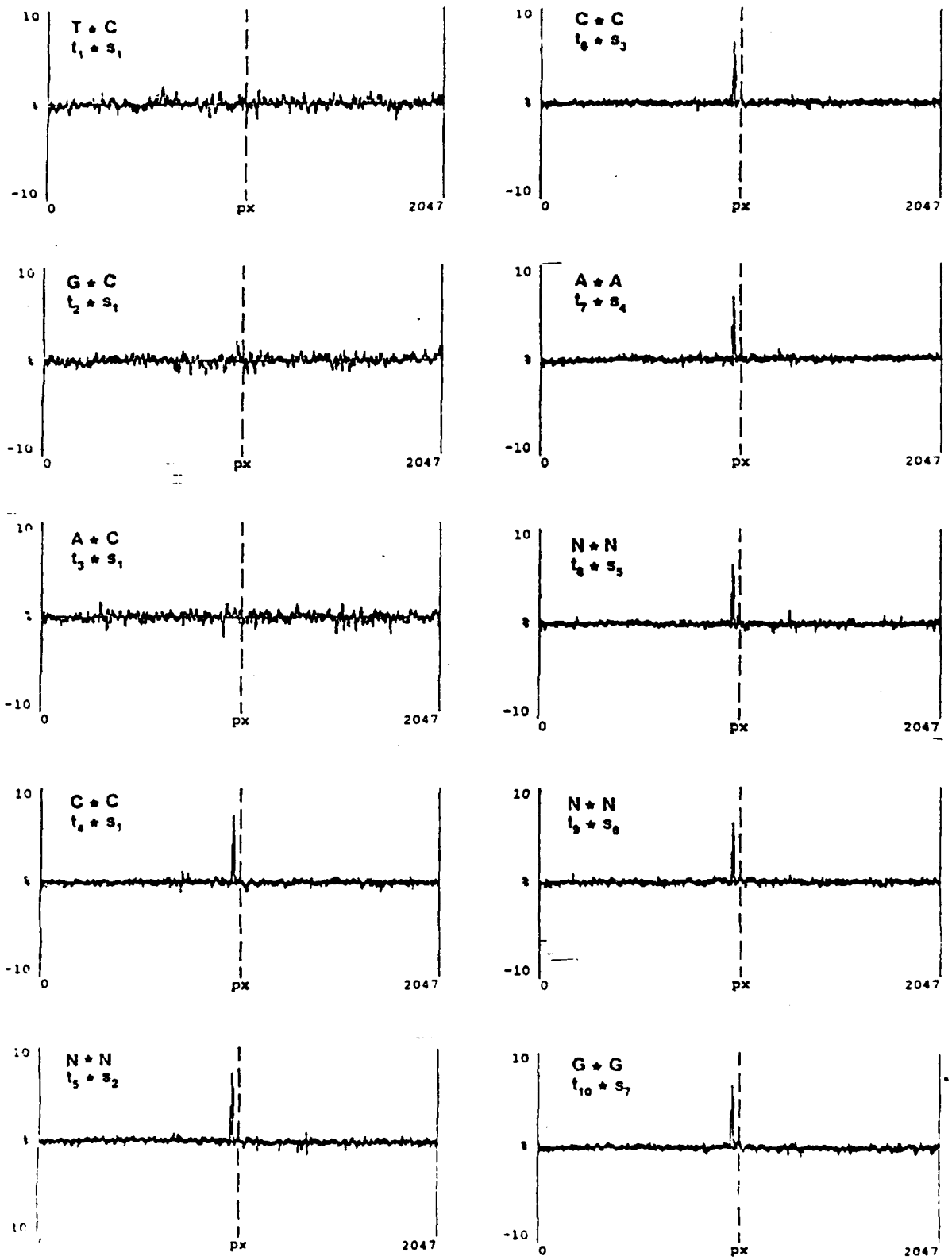


Figure 17: Correlations produced by a fine analysis performed with a 7-bases query sequence that is identical to a segment of a 20-bases long database.

segment of the database containing the query sequence. The following bases of the query sequence, t5, t6, t7, t8, t9 and t10 were respectively correlated with the bases s2, s3, s4, s5, s6 and s7 of the database. The positions t4, t5, t7, t8 and t10 were occupied by identical bases but the positions t6 and t9 had different bases. This time, it was concluded that the query sequence was not identical to the segment t4-t10 of the database but only similar. Figure 18 shows the ten correlation patterns obtained during that experiment. A more elaborate procedure could determine the identity of the bases of the database that do not match the query sequence. In such a procedure, each time that a discrepancy was found between the query sequence and the database, three supplementary trials would be required to identify the base of the database that is different from the query sequence.

6.0 CONCLUSION

Elements of optical data processing and spread-spectrum communication theory have been integrated to design and demonstrate the analysis of DNA sequences with an optical TIC. An analysis strategy including a coarse and a fine analysis was developed and the resulting processing times were calculated. It was concluded that TICs could produce a substantial improvement in DNA analysis processing times. Comparison of the processing time for a particular case lead to the conclusion that the TIC is 10 times faster than a 80 MIPS computer and over 375 times faster than a personal computer.

The feasibility of the coarse and the fine analysis was demonstrated experimentally and the capability of a TIC to produce a correlation peak with as much as a 50% discrepancy between the query sequence and the corresponding segment of the database was demonstrated. The requirements of an operational system were outlined.

7.0 ACKNOWLEDGEMENT

The authors gratefully thank John Bodie and Marion Power from Calian Communications Systems Ltd. for their assistance in circumventing the Controller idiosyncrasies. The assistance of Lt. S. Faulkner and Dr. A. Gulliver with coding theory was greatly appreciated.

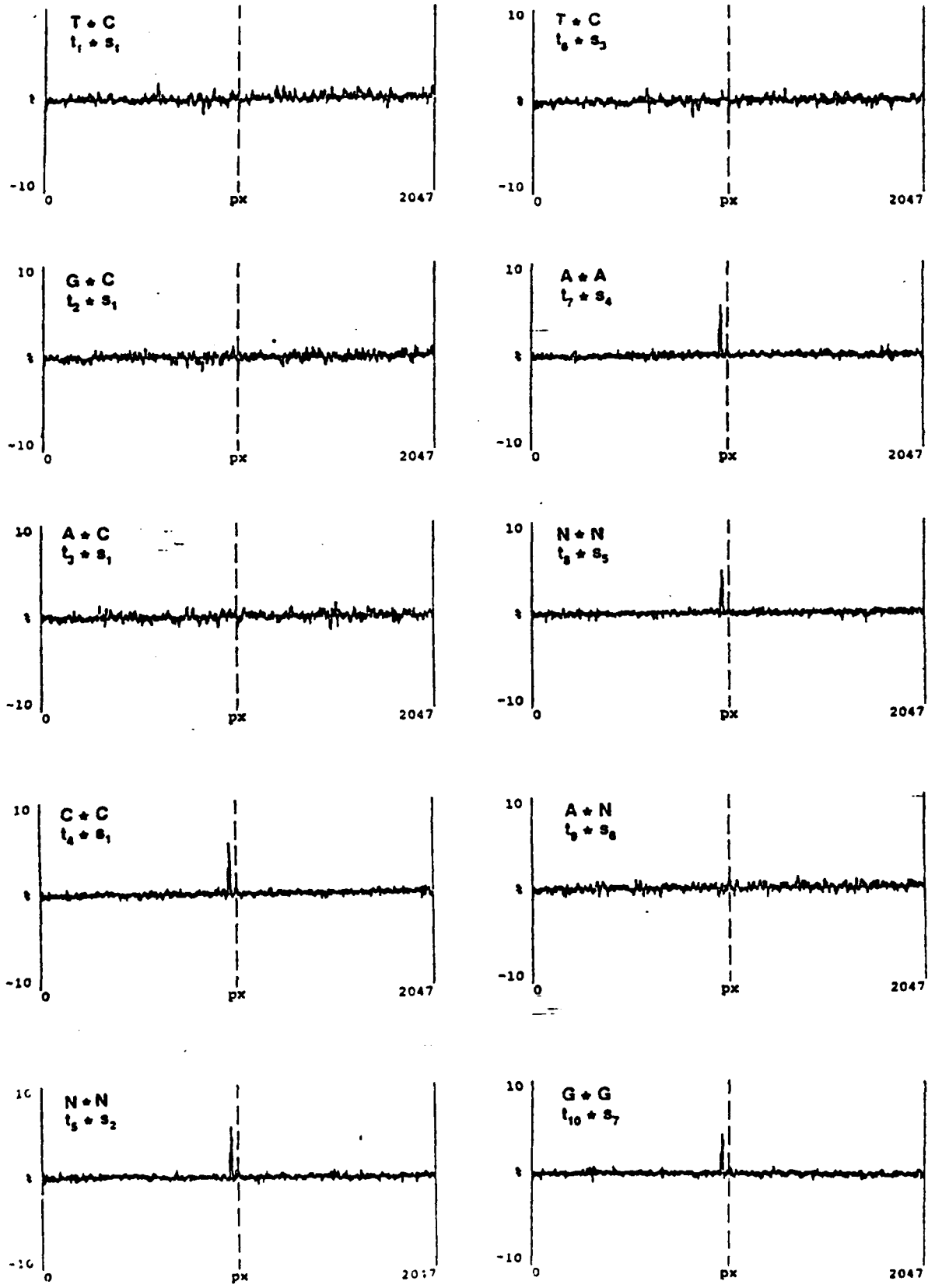


Figure 18: Correlations produced by a fine analysis performed with a 7-bases query sequence that is similar to a segment of a 20-bases long database.

8.0 REFERENCES

[1] L. Smith and L. Hood, "Mapping and Sequencing the Human Genome: How to Proceed", BIO/TECHNOLOGY vol.5, Sept. 1987, p.933-939.

[2] R. Lewis,"How Lasers Can Speed Up The Human Genome Project", Photonics Spectra, May 1991, p.72-75.

[3] S.L. Williams, "Imaging the Human Genome", Advanced Imaging, July 1990, p.16-19.

[4] N. Brousseau and R. Brousseau, 'Analysis of DNA Sequences with an Optical Time-Integrating Correlator: Proposal", DREO TN.

[5] R.C. Dixon, "Spread Spectrum Systems", John Wiley & Sons, 1984.

[6] S.W. Golomb, "Shift Register Sequences", Aegean Park Press, Revised Edition 1982.

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (the name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Establishment sponsoring a contractor's report, or tasking agency, are entered in section 8.)

NATIONAL DEFENCE
DEFENCE RESEARCH ESTABLISHMENT OTTAWA
SHIRLEY BAY, OTTAWA, ONTARIO K1A 0K2 CANADA

2. SECURITY CLASSIFICATION (overall security classification of the document including special warning terms if applicable)

UNCLASSIFIED

3. TITLE (the complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S,C or U) in parentheses after the title.)

ANALYSIS OF DNA SEQUECES WITH AN OPTICAL TIME-INTERGRATING CORRELATOR: PROOF-OF-CONCEPT EXPERIMENTS (U)

4. AUTHORS (Last name, first name, middle initial)

BROUSSEAU, N., SALT, J.W.A., GUTZ, L. AND TUCKER, M.D.B.

5. DATE OF PUBLICATION (month and year of publication of document)

MAY 1992

6a. NO. OF PAGES (total containing information. Include Annexes, Appendices, etc.)

41

6b. NO. OF REFS (total cited in document)

6

7. DESCRIPTIVE NOTES (the category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

DREO TECHNICAL NOTE

8. SPONSORING ACTIVITY (the name of the department project office or laboratory sponsoring the research and development. Include the address.)

NATIONAL DEFENCE
DEFENCE RESEARCH ESTABLISHMENT OTTAWA
SHIRLEY BAY, OTTAWA, ONTARIO K1A 0K2 CANADA

9a. PROJECT OR GRANT NO. (if appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant)

041LQ

9b. CONTRACT NO. (if appropriate, the applicable number under which the document was written)

10a. ORIGINATOR'S DOCUMENT NUMBER (the official document number by which the document is identified by the originating activity. This number must be unique to this document.)

DREO TECHNICAL NOTE 92-12

10b. OTHER DOCUMENT NOS. (Any other numbers which may be assigned this document either by the originator or by the sponsor)

11. DOCUMENT AVAILABILITY (any limitations on further dissemination of the document, other than those imposed by security classification)

(X) Unlimited distribution
( ) Distribution limited to defence departments and defence contractors; further distribution only as approved
( ) Distribution limited to defence departments and Canadian defence contractors; further distribution only as approved
( ) Distribution limited to government departments and agencies; further distribution only as approved
( ) Distribution limited to defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availabilty (11). However, where further distribution (beyond the audience specified in 11) is possible, a wider announcement audience may be selected.)

UNCLASSIFIED

SECURITY CLASSIFICATION OF FORM

13. ABSTRACT (a brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), or (U). It is not necessary to include here abstracts in both offical languages unless the text is bilingual).

(U) The analysis of the molecular structure called DNA is of particular interest for the understanding of the basic processes governing life. Correlation techniques implemented on digital computers are currently used to perform the analysis but the present process is so slow that the mapping and sequencing of the entire human genome requires a computational breakthrough. This paper presents proof-of-concept experiments of a new method of performing the analysis of DNA sequences with an optical time-integrating correlator. Included are experimental results for the two types of analysis specified by the processing strategy. Details about the design and construction of the custom signal generators that were built to perform the experiments are presented.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus. e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus-identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

DNA ANALYSIS
TIME-INTEGRATING CORRELATOR
OPTICAL DATA PROCESSING

UNCLASSIFIED

SECURITY CLASSIFICATION OF FORM