

Mining DNA Data in an Efficient 2D Optical Architecture

Jason M. Kinser
George Mason University
Manassas, VA 20110
jkinser@gmu.edu

ABSTRACT

Optical computing is applicable to very large database problems that require several iterations of processing. One such forthcoming problem is the analysis of DNA data, in which an individual in a population may have billions of bases in their DNA structure. Current methods only find simple first order associations between the DNA structure and the expressed illnesses. This means that the presence or absence of a gene strongly influences the onset of an illness. These current methods will have a difficult time in finding higher order associations between the genome and the expressed phenotypes. The discovery of such associations will take several iterations through a large database to extract association information. Searching a database of billions of elements for each of hundreds of individuals through thousands of iterations is prohibitively expensive on electronic computers. Highly parallel architectures such as optical correlators will offer the opportunity to process this large amount of data in a higher-order fashion.

Keywords: DNA, optical correlation.

1. INTRODUCTION

Bioinformatics is a burgeoning field that is developing larger information processing requirements as the size of the accumulated databases rapidly increase. Optical correlators have the potential of rapidly searching large databases in a massively parallel fashion. Thus, it becomes prudent to explore the marriage of optical correlators and bioinformatic problems.

Bioinformatics encompasses several sub-fields. For this discussion one such field will be pursued. This will be the problem of finding target DNA strings within a database. There is one major concern with the implementation of DNA data into an optical system. That concern is how to encode the DNA information into an optical image. DNA data is one-dimensional and the nucleotides (C, A, G, and T) are independent of each other. Whereas, optical systems become far more efficient when the information is presented as an image. Initially, it appears as though DNA information and optical processing are ill-matched.

This document will explore a method of encoding DNA data suitable for optical processing, provide a simple architecture, and present results from simulations.

2. ENCODING DATA

DNA information is usually reduced to a string of letters. The alphabet contains four elements (C, A, G, and T) to represent the four constituents of a DNA strand. It is this string that needs to be encoded.

2.1. Independence

The four members (nucleotides) of the alphabet are independent. Thus, it is necessary to avoid encoding schemes that violate this independence. For example, an encoding in which each nucleotide is assigned an integer (e.g., C=1, A=2, G=3, and T=4) can not be used. This system implies such absurdities as $C+A=G$ and $T>G$, etc. These implications do not exist in the data and must be avoided, if possible.

2.2. Raster Imaging

To realize the full parallel ability of an optical system the data should be represented as a two-dimensional image in which the elements may be complex numbers. Of course, this is not in congruence with the one-dimensional nature of the DNA data. Furthermore, either dimension of a realizable two-dimensional image will have far fewer elements than the length of the DNA strand.

Another complication is that the order of the DNA strand must be maintained. Thus, presenting a DNA strand as a raster image causes problems in that neighboring nucleotides are separated and non-neighboring nucleotides become neighbors. This problem is display in figure 1.

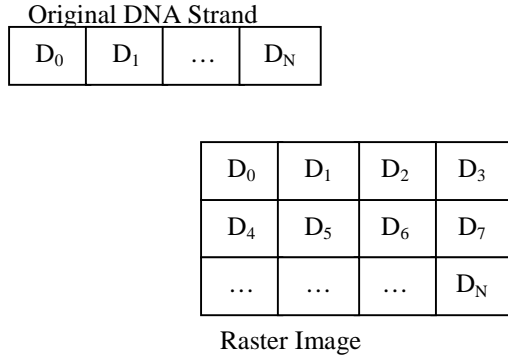


Figure 1. The Encoding of a DNA Strand as a Raster Image.

In this case, it is easy to see that elements D_3 and D_4 which were originally neighbors are now separated and elements such as D_3 and D_7 which are not originally neighbors are now such. This type of encoding must be avoided because the spatial relationship between the elements is distorted.

2.3. Requirements

Thus, the two requirements for encoding the DNA data stream is to maintain the independence of the elements and to not disturb the spatial relationship.

2.4. Solution 1. Spectrograms.

The first solution attempted borrows from the field of speech processing the idea of a spectrogram. Speech signals are long one-dimensional signals in which the spatial relationships must be maintained. Often speech data is represented as a spectrogram using short-window Fourier transforms such as,

$$M_{ij} = \mathfrak{S}\{G_j\{S\}\}_i \quad (1)$$

where

$$G_j = S[j * m : j * m + n] \quad (2)$$

and $m \approx n/k$.

The variable n is the size of a short window (also equivalent to twice the vertical dimension of M). The variable m is the distance that the window is shifted for each j and k is greater than 1 (often in the range from 3 to 10). Basically, G_j is a sample of length n . The G_{j+1} will overlap G_j by an amount determined by k .

Before we can create the spectrogram, we still have to put the data into a numerical form. It is also desired that the amount of data not increase by a factor of four during this encoding process. Here we

rely on a natural grouping of the nucleotides. Purines are A and G and pyrimadines are C and T. We can make A and C orthogonal my $C=1$ and $A=1j$. The negatives in this group are then $G=1j$ and $T=-1$. While this isn't complete orthogonality, this encoding does alleviate some of the more serious concerns without increasing the amount of data.

The spectrogram of a DNA string thus encoded is shown in figure 2.

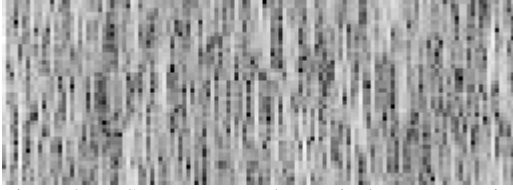


Figure 2. A Spectrogram. The vertical component is frequency and the horizontal component is position in the data stream.

This method maintains the relative position of the data. However, it is not very efficient. The number of pixels is $V*H$ (V =vertical number and H =horizontal number) and the number of elements considered from $V+m*H$. To increase the number of pixels would sacrifice the amount of overlap in the short windows. This will cause problems in an effort to recognize data strings that exceed a short window boundary. In this method there is a trade-off between pixel efficiency and continuity.

2.5. Simple Test for Solution 1.

Here the use of the previous encoding scheme is examined in a simple example. The goal is to recognize a subsequence of a larger data stream. So, the database was encoded in a spectrogram and a DNA segment was similarly encoded. The two images were then correlated to find similarities. The correlation surface is shown in figure 3 and the plot of the horizontal slice of the correlation surface is shown in figure 4.

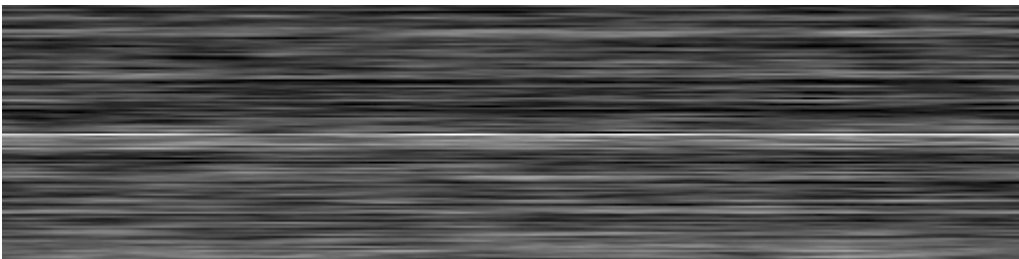


Figure 3. The correlation surface.

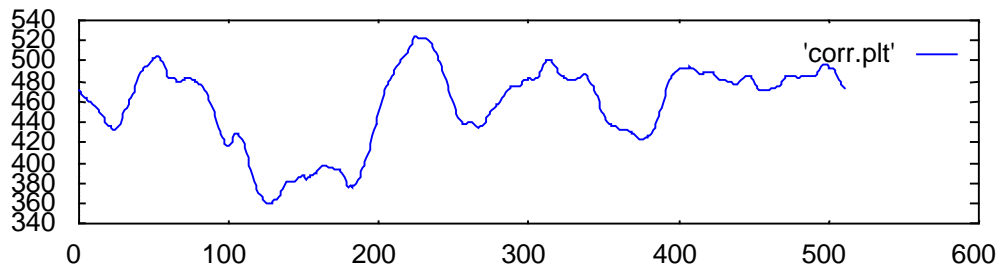


Figure 4. The plot through the middle of the correlation surface.

Obviously, the correlation surface does not display a single peak that would indicate a match between the database and the segment. The database contained 8304 nucleotides and the extracted segment started at 1586. So, we would expect to see a peak at $0.19*H$ in the correlation image. No such peak exists.

Other trials displayed similarly dismal results. So, this encoding method is no longer deemed as a viable option.

2.6. Solution 2.

This solution relies on the optical property of diffraction. Consider the diffraction architecture in figure 5.

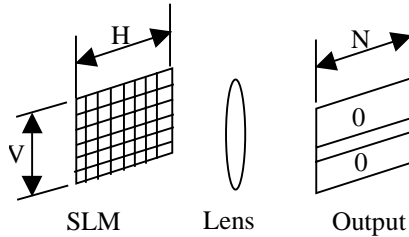


Figure 5. A simple architecture to consider the effects of diffraction.

It is the purpose of this architecture to use the $V \cdot H$ pixels in an SLM to create a N element diffraction pattern that is one-dimensional in the output plane. This becomes effective if N is much larger than H . The purpose is to use diffraction to create the one-dimensional signal of the DNA stream.

Here there are competing attributes in the system. The number of pixels in the SLM is smaller than the number of elements in the output space. Normally, this would be unacceptable since the number of degrees of freedom in the input is less than the variables in the output and this is a first order system. However, optical computations are not very precise and small errors can be traded for global optimization.

If this design were to be used for an optical correlator then the output would be in Fourier space. So, the system would try to produce the Fourier transform of a complex one-dimensional signal in the output space.

The main concern is the ability to increase the dimensionality in the output space. Is it possible to increase the spatial bandwidth in the horizontal dimension? In order to accomplish this a random phase mask (of the same spatial bandwidth as the SLM) is placed behind the SLM. This will give the elements in the SLM the opportunity to interfere and create a higher resolution signal in the output plane.

A simulated annealing program was employed to generate an SLM image that would create a signal similar to a target. The target image was created by encoding a DNA stream into a vector of complex numbers and then obtaining the Fourier transform of that sequence. In this simulation the amplitude and phase of each element of the SLM were coupled. Thus, an increase in input voltage would increase both amplitude and voltage. The illumination from a single pixel with input voltage v was,

$$\text{illumination} = v \exp[j2\pi v] \quad (3)$$

where v ranged from 0 to 1. In this fashion, the simulation represented the coupling restriction of SLM elements.

Typical results are shown in figure 6 and 7. Figure 6 displays the output image and the center row is the data segment. Figure 7 plots this row against the original target signal. In this example, the horizontal dimension was increased by a factor of two. The SLM was 32×32 .

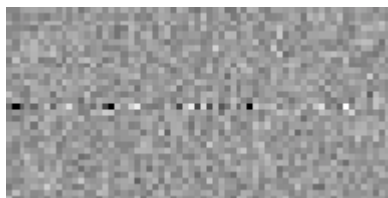


Figure 6. A created output from an SLM. The center row contains the data and the rest of the pixels are close to 0.

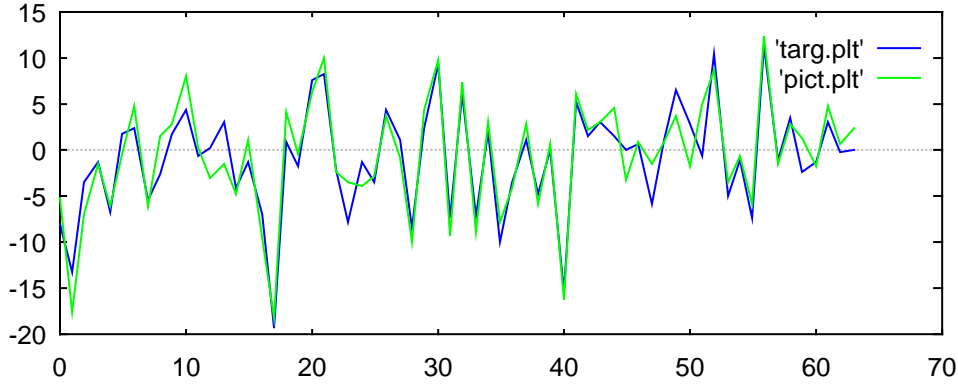


Figure 7. The target and created (pict) Fourier space arrays.

While there is some error in the recreation of the target, they are relatively small. Figure 8 displays an eight-fold expansion and as seen with this method the errors increase.

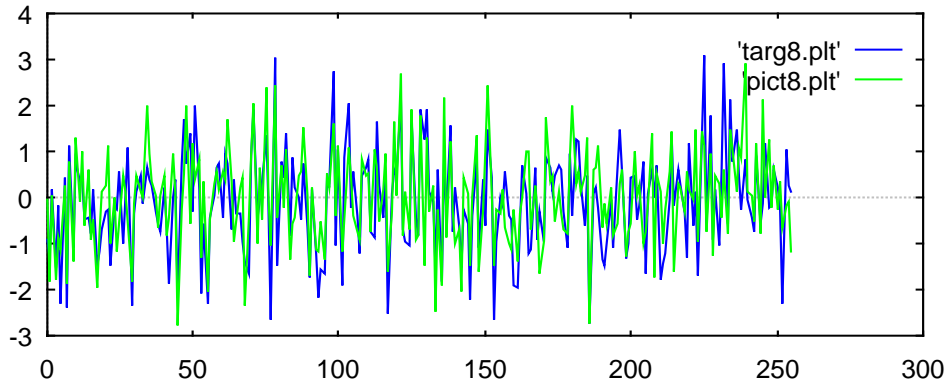


Figure 8. An eight-fold expansion.

2.4. Simple Test for the Second Solution.

More important than perfect re-creation is the ability to recognize a DNA strand within a database. In the current configuration $M=2$ and $H=32$. Thus, there are 64 nucleotides that have been encoded. To test the ability to correlate and a new signal was created from a 10 nucleotide segment of the original 64 nucleotides.

This new sequence was converted to the closest SLM image through the same process as in 2.3. Now, there are two SLM arrays that create two Fourier images. The first attempts to recreate the Fourier representation of the entire 64-nucleotide database. The second attempts to recreate the Fourier representation of the 10-nucleotide segment.

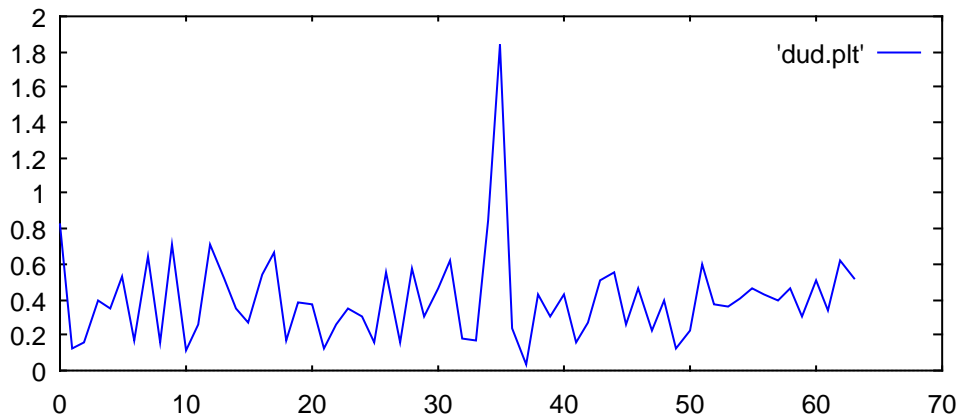


Figure 9. Correlation surface between a 10 element segment and a 64 element database.

The plot in figure 9 depicts the correlation surface between these two generated vectors. The large spike indicates that the segment was located within the database. Thus, the small infractions received from the SLM recreation of the Fourier signal are not significantly deteriorating the correlation.

3. OPTICAL ARCHITECTURE

The preliminary tests indicate that such a design is possible. This section will present an optical design.

3.1. The Foundation

The purpose of this optical engine is to compute the existence of segments within a database. It would be useful to examine more than one database at a time. Thus, the plan is that several linear databases will be stored on the filter plane and the SLM would create several copies of the search target. Each copy would then be optically correlated with the database. This design is shown in figure 10. In this system the segment input is written to the SLM and it creates several horizontal copies of the target. Since each copy is alike there will be vertical stripes in this image. The filter is stored on film, a hologram, or another SLM. This image contains the Fourier components of several DNA files. Each of these is stored as a horizontal array. The correlations of each segment with its respective target would pass through a cylindrical lens that would perform the Fourier transform in one dimension. The output would be a set of one dimensional correlation surfaces that would then indicate the presence and location of the segment within each vector in the database.

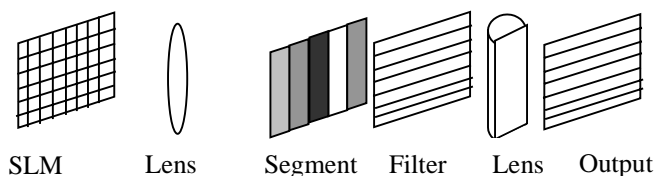


Figure 10. The optical architecture.

3.2. Simulation.

There are a couple of differences from this design than from the previous tests. The first is that the database is now stored without consideration as to the representation imposed by the SLM. In other words, it does not contain the errors encountered earlier. The second item is that the image produced from the SLM will now be required to be vertically homogeneous instead of having only a horizontal vector of information.

The same simulated annealing program was employed to produce an SLM image that was vertically uniform. An example of such an image from a two-fold expansion experiment is shown in figure 11. Now, each horizontal row of this image contains the Fourier space representation of the input DNA stream. Each of these will independently correlate with the data stored at the filter plane.

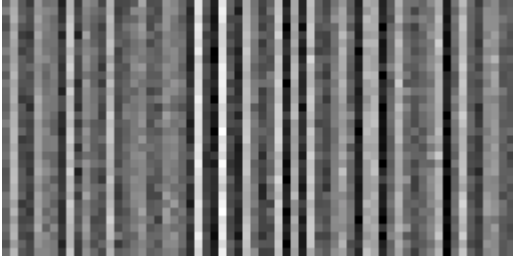


Figure 11. An output from an SLM attempting to achieve vertical uniformity.

In an experiment a 10 nucleotide segment was removed from a data stream. The filter was constructed from several overlapping segments of the same data stream. These segments are 64 elements in length. The correlation of this a vector in the filter plane and the input from figure 11 is shown in figure 12. As can be seen the correlation spike is quite distinct. Recall that the input contained SLM imposed errors while the stored filter did not.

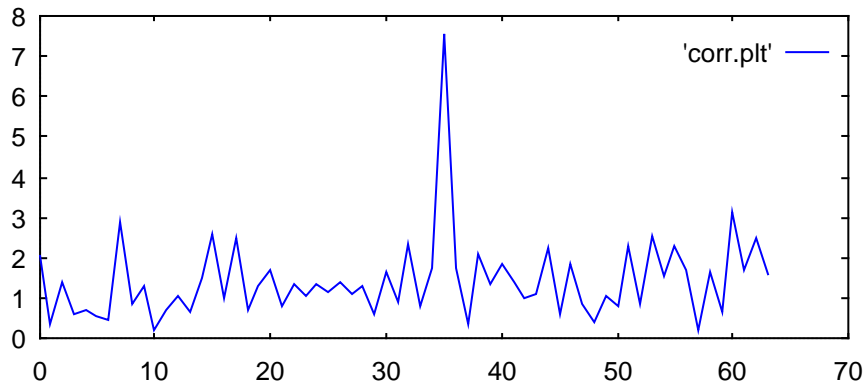


Figure 12. The correlation from on row of the SLM generated image and one segment in the filter that contains the target.

Figure 13 displays the output correlation image from all the correlations. Each successive row of the filter overlapped the previous row by 54 elements. Thus, in this experiment it is expected that the correlation spike will shift by 10 pixels for subsequent rows of the output. This is clearly seen in figure 13. The top row of this image corresponds to the plot in figure 12.



Figure 13. Correlation output of the optical simulation.

Thus, it is possible to find the target within the database using this method.

4. FUTURE ENDEAVORS

There is still much work that needs to be done with this problem. For example, the use of a random phase mask instead of one optimally designed should be considered. A second question that remains concerns the relationship between the induced SLM error and the factor m .

ACKNOWLEDGMENTS

This work was supported on a DARPA contract through Starzent Inc.