Mining DNA Sequences Based on Spatially Coded Technique Using Spatial Light Modulator

Masomee Fazelian, Sajjad AbdollahRamezani, Sima Bahrami, Ate Chizari, Mohammad Vahid Jamalil, Pirazh Khorraramshahi, Arvin Tashakori, Shadi Shahsavari, and Jawad A. Salehi, Fellow, IEEE
Optical Networks Research Laboratory (ONRL), Sharif University of Technology (SUT), Tehran, IRAN

Index Terms—String data alignment, moiré pattern, DNA sequencing, spatial light modulator.

I. INTRODUCTION

Emerging various widespread human diseases speeds up the growing rate of genomics. Accordingly, analysis of deoxyribonucleic acid (DNA) sequences, as a medium storing genetic information of the genome and the expressed phenotypes could more effectively be extracted. Such moiré fringes reveal occurrence of matching, deletion and insertion between DNA sequences providing useful visualized information for prediction of gene function and classification of species. Furthermore, by applying a cylindrical lens, a new technique is proposed to map two-dimensional (2D) association information to a one-dimensional (1D) column of pixels, where each pixel in the column is representative of superposition of all bright and dark pixels in the corresponding row. By such a time-consuming preprocessing, local similarities between two intended patterns can readily be found by just using a 1D array of photodetectors and post-processing could be performed on specified parts in the initial 2D pattern. We also evaluate our proposed circular encoding adapted processing could be performed on specified parts in the initial 2D pattern. We also evaluate our proposed circular encoding adapted processing could be performed on specified parts in the initial 2D pattern. We also evaluate our proposed circular encoding adapted

II. PRINCIPLES

In this section, the principles of string alignment by moiré technique are outlined. Consider two data sequences. The goal of string alignment is evaluation of similarities and differences between them. In particular, we are interested in distinguishing insertion and deletion of elements in any strings with respect to each other. Moiré technique applies high speed parallel processing of light to perform string alignment. In this approach, four components of strings, namely \{A, G, C, T\} are encoded as \{1000, 0100, 0010, 0001\}, respectively. Based on this coding, the strings are spatially coded into images where each component corresponds to four narrow stripes with one bright stripe as “1” and three dark stripes as “0” (see Fig. 1). The coded images are then overlapped with a small relative
angle, and by using this technique, correlating segments of the
second string in various shifts of the first one can evidently
be distinguished. The subsequent matched elements will be
appears as a bright line in the observed pattern of overlapped
images.

As an example, consider two strings $S_1$ of length 40 and $S_2$ of length 20. Now, we want to search for $S_2$ in $S_1$. Fig. 2(a) shows $S_2 = \{ACGTATCGTGACGTCGAA\}$ with respect to the codes appeared in Fig. 1 and each row in Fig. 2(b) shows subsequent shifts of initial string $S_1 = \{TCCGTATCTCCGTACAGGTCGAATGCGTACATCGACCT\}$; for example first row shows $S_1(1:20)$, second row shows $S_1(2:21)$, up to the last row. Overlapping Fig. 2(a) ans (b) results in the pattern shown in Fig. 2(c); the bright line in the fourth row illustrates that a correlation has happened for a shift of 6, i.e., $S_2$ and $S_1(6:25)$ are matched.

The insertion and deletion of elements lead to a vertical
shift in some parts of the bright line in the overlapping
pattern. Each break point indicates the location where
insertion or deletion is occurred. The positive and negative
vertical shifts correspond to insertion and deletion of some
elements, respectively. As an example, consider the string
$S_3 = \{ACGTATAGCCGTACAGGTCGAA\}$ generated by
insertion of "AG" between the sixth and seventh element
of $S_2$ and deletion of the fourteenth and fifteenth element
of $S_2$. Figure 2(d) depicts the output pattern obtained by
multiplying $S_3$ and Fig. 2(b).

### III. PROPOSED METHODS

In this section, we propose several practically feasible moiré
patterns for string data alignment applications. Wave nature
of light provides enough degrees of freedom, i.e., amplitude,
phase, and polarization manipulation for sequence data pro-
cessing.

The first coding approach is based on correlation in which
a sequence is simply symbol-by-symbol compared to another
sequence. In DNA sequence data processing, each symbol
denotes a DNA base. In comparing two symbols with each
other, similar symbols generate a bright spot; hence, a cor-
related set realizes a bright line. This line is fragmented in
the case of insertion and deletion in which vertical distance
between fragmented lines identifies the number of deleted
or inserted elements in that place. In this method, SNR
can easily be calculated; in the case of two independent
and identical distributed sequences, the probability of such
a random similarity and hence number of bright spots with
respect to full matching is $0.25$, leading to $6 \text{ dB SNR}$. It is
notable that system’s SNR is proportional to the ratio of bright
line intensity to the average intensity of other rows.

Another coding technique is based on concatenating two
subsequent elements, for example $S(i : i + 1)$ and $S(i + 1 : i + 2)$, as a group. Subsequent groups have a common element
which ensures an easier detection procedure of insertion or
deletion. Coding sequences in overlapped pairs not only does
increase the SNR but also makes correlated elements more
distinguishable even in the cases of insertion and deletion.
In this method, a $12 \text{ dB SNR}$ can be expected in that the
probability of random similarity for a word of two symbols is
$0.0625$. 

| Table I: Corresponding codes for polarized spatial patterns in Figs. 3 and 4. |
|---------------------------------|---|---|---|---|
| DNA bases | A | G | C | T |
| Type I    | 1000 | 0100 | 0010 | 0001 |
| Type II   | H00H | V0V0 | 01V0 | 00H0 |

$H$: Horizontal, $V$: Vertical
TABLE III

|                  | Type I                  | Type II                  | Type III                | Type IV                 |
|------------------|-------------------------|--------------------------|-------------------------|-------------------------|
| **Processing Gain** | $N/4$                  | $N/4$                     | $N/8$                   | $N/16$                  |
| **SLM Modulation Capacity** | Intensity or Polarization | Intensity and Polarization | Intensity and Polarization | Intensity or Polarization |
| **SNR (dB)**      | 6.8854                  | 6.4648                    | 12.2250                 | 12.0715                 |

A. Bar Pattern

We examined two different sets of symbols in a bar moiré pattern. While the first one employs pulse position modulation (PPM), the second comprises of a set of four orthogonal codes using both intensity and polarization (see Table I). Since there is no useful information in shifts that are not an integer product of symbol length, different rows are shifted by an integer product of four slots that form a symbol. This is by far more efficient than horizontal tilting of the second pattern and consequently compatible with finite resolution of SLM.

Simulation results are depicted in Figs. 3 and 4. By comparing the results, it is clear that using type II increases the intensity of both noise and signal but does not improve the SNR. Measuring symbol-by-symbol correlation, we see the SNR does not go further than 6 dB.

In the second approach, the codes in Table II are applied which means that for tilted pattern different rows are shifted by $8k$ in type III (word length is eight here) and $16k$ in type IV; $k$ is a positive integer. Figs. 5 and 6 illustrate the simulation results. As it can be seen, the horizontal straight line is more vivid in types III and IV since the probability of random similarity for a word of two symbols is 0.0625; therefore, we can expect a SNR about 12 dB. In type III, we need a SLM with independent intensity and polarization modulation while in type IV only intensity or polarization modulation is required. Polarization modulation can easily be converted to intensity modulation via a polarizer. Since word length for type IV is twice type III, for an equal number of SLM surface pixels, processing gain, the number of DNA bases that the setup is able to compare in each run, of type III is twice type IV. Types III and IV offer better detection capability facing insertion and deletion. It is notable that each insertion or deletion changes two words. In case of $n$ subsequent deletion or insertion, $n + 1$ words differ from initial pattern. Moreover, if we increase the word length to code the DNA bases in group of length $L$, $L + 1$ elements are needed to be overlapped in order to detect insertion and deletion. When misalignment and other types of errors are addressed, the maximum performance of such a system could be achieved. In this case, the number of pixels of SLMs to compare two sequences of length $N$ follows Table III. It also reports the SNR values of different types for a random sequence of length 48.

B. Circular Pattern

Optical alignment could be quite problematic in implementing bar patterns. In correlating two bar patterns, the dimension precision required should be about $d/N$, where $d$ is the transverse length of a pixel and $N$ is the total number of vertical pixels on SLM surface. On the other hand, circular moiré patterns are basically easier to be adjusted in experimental setups since only the center of circles should be aligned. Besides, it is sensitive to neither rotation nor
From the above equation, \( \Delta r_1 \) is chosen such that 
\[
\Delta r_1 = \Delta \theta r_0 = r_1 \Delta \theta \Delta r_0.
\] (1)

Hence, we can define the following recursive relation to obtain 
\[ r_{i-1} \Delta r_{i-1} \] respectively. Fig. 8(c) depicts the simulation and the experimental patterns after overlapping two images of Figs. 8(a) and (b). As can be seen, a bright ring appears at the intersection of matched elements.

IV. EXPERIMENTAL SETUP AND RESULTS

In this paper, the optical architecture is implemented by two separate programmable reflective SLMs. The pixel pitch of the liquid-crystal display of each SLM is 20 \( \mu \)m, and the pixels number is 1280 \times 768. The larger number of elements it comprises, the less resolution in the output plane we achieve. Further details about equipments can be found in Table III. Fig. 9 shows the architecture of the proposed optical processing method for genome analysis based on spatial coded technique. The output light of a spatially coherent source such as a laser diode source or a non-coherent source like LED is collimated and then impinges the first SLM containing \( S_1 \). A linear polarizer in front of the first SLM sets the incoming polarization state. Since the laser emits elliptical polarized light, the intensity is dependent on both the SLM and the first polarizer’s states. To remove this ambiguity, the intensity of the light leaving the first polarizer has to be set to be independent of the angle.

The reflected light from the first SLM meets the second SLM which implements \( S_2 \). The output light of a spatially coded technique. The output light of a spatially coded technique. The output light of a spatially coded technique. The output light of a spatially coded technique. The output light of a spatially coded technique. The output light of a spatially coded technique. The output light of a spatially coded technique.

In order to verify our proposed method, we firstly show bar strings alignment between two DNA-simulated sequences; then the circular one will be demonstrated as our proposed new encoded pattern. One-dimensional strings to be aligned are illustrated in Figs. 10 and 11 in which \( S_1 \), \( S_2 \), and \( S_3 \) were introduced earlier. To more straightforwardly realize string
alignment, a cylindrical lens could be employed between the third polarizer and the output display. It is well known that such a lens transforms plane wave to an ultra-thin line. As a result, each horizontal bright line in the output pattern right behind the lens is mapped to a luminous point on the display which enables us to use a simple one-dimensional array of photodetectors to detect the occurrence of exact matching and the number of deleted or inserted elements. Figs. 12(a) and 12(b) respectively illustrate the transformed versions of the output patterns in Figs. 10(f) and 11(f) at the focused plane of the cylindrical lens. Additionally, simulated and experimental results for circular patterns presented in Fig. 8 are in good agreement.

V. Conclusion

In conclusion, a simple and practical method based on spatially coded moiré matching technique has been proposed for string alignment processing. Easy interpretation and inherent parallelism with almost real-time processing are the main specifications of our approach which is compatible with digital devices. The processing gain and SNR of the proposed patterns, i.e., bar and circular patterns, have numerically been calculated to show the effectiveness of our method. Moreover, a preprocessing stage which remarkably decreases post-processing time needed for interpretation of output pattern has been introduced. The capability of our proposed method in DNA sequence matching has been shown via simulation. Finally, experimental results verify the performance of the method in genomics processing applications based on optical computing.
REFERENCES

[1] J. M. Kinser, “Mining DNA data in an efficient 2d optical architecture,” in 2000 International Topical Meeting on Optics in Computing (OC2000). International Society for Optics and Photonics, 2000, pp. 104–110.

[2] J. Shendure and H. Ji, “Next-generation DNA sequencing,” Nature biotechnology, vol. 26, no. 10, pp. 1135–1145, 2008.

[3] A. C. Rajan, M. R. Rezapour, J. Yun, Y. Cho, W. J. Cho, S. K. Min, G. Lee, and K. S. Kim, “Two dimensional molecular electronics spectroscopy for molecular fingerprinting, DNA sequencing, and cancerous DNA recognition,” ACS nano, vol. 8, no. 2, pp. 1827–1833, 2014.

[4] J. Erd, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman et al., “Real-time DNA sequencing from single polymerase molecules,” Science, vol. 323, no. 5910, pp. 133–138, 2009.

[5] J. M. Rothberg, W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards et al., “An integrated semiconductor device enabling non-optical genome sequencing,” Nature, vol. 475, no. 7356, pp. 348–352, 2011.

[6] S. K. Min, W. Y. Kim, Y. Cho, and K. S. Kim, “Fast DNA sequencing with a graphene-based nanochannel device,” Nature nanotechnology, vol. 6, no. 3, pp. 162–165, 2011.

[7] M. L. Metzker, “Sequencing technologies: The next generation,” Nature reviews genetics, vol. 11, no. 1, pp. 31–46, 2010.

[8] Z. Ning, A. J. Cox, and J. C. Mullikin, “Ssaha: a fast search method for large dna databases,” Genome research, vol. 11, no. 10, pp. 1725–1729, 2001.

[9] W. J. Kent, “Blat: the blast-like alignment tool,” Genome research, vol. 12, no. 4, pp. 656–664, 2002.

[10] J. Tanida, “String data alignment by a spatial coding and moiré technique,” Optics letters, vol. 24, no. 23, pp. 1681–1683, 1999.

[11] J. Tanida and K. Nitta, “String data matching based on a moiré technique using 1d spatial coded patterns,” in 2000 International Topical Meeting on Optics in Computing (OC2000). International Society for Optics and Photonics, 2000, pp. 16–23.

[12] E. R. Mardis, “Next-generation DNA sequencing methods,” Annu. Rev. Genomics Hum. Genet., vol. 9, pp. 387–402, 2008.

[13] J. L. Merkling, “Sequence matching in holographically stored genetic strings,” Ph.D. dissertation, Texas Tech University, 2005.

[14] J. Tanida, K. Nitta, and A. Yahata, “Spatially coded moiré matching technique for genome information visualization,” in Photonics Asia 2002. International Society for Optics and Photonics, 2002, pp. 26–33.

[15] K. Nitta, H. Togo, A. Yahata, and J. Tanida, “Genome information analysis using spatial coded moiré technique,” in Lasers and Electro-Optics, 2001. CLEO/Pacific Rim 2001. The 4th Pacific Rim Conference on, vol. 2. IEEE, 2001, pp. II–II.

[16] I. Amidror, The theory of the moiré phenomenon. Springer, 2000, no. LSP-BOOK-2000-001.

[17] E. Gabrielyan, “The basics of line moiré patterns and optical speedup,” arXiv preprint physics/0701068, 2007.