String data alignment by a spatial coding and moiré technique

Jun Tanida

Department of Material and Life Science, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita 565-0871, Japan

Received August 18, 1999

A method for string alignment is presented in which a moiré technique is applied to one-dimensional spatial encoding patterns. String alignment, an essential operation in genome analysis, evaluates local similarity between sequences of bases or amino acids. The method uses a simple procedure to provide matching results for not only the same locus but also neighboring loci. Experimental verification shows the effectiveness of the proposed method. © 1999 Optical Society of America

OCIS codes: 200.0200, 200.3050, 200.4560, 000.1430, 000.4920.

Optical computing techniques that use the physical properties of light are expected to provide effective means for information processing with the help of advanced optoelectronic technologies. In view of the past and current status of optical computing research,[1] at least four requirements should be satisfied for practical application of these techniques, i.e., establishment of fast throughput of data, avoidance of precise adjustment, effective use of optical properties, and selection of appropriate applications. Because current optoelectronic devices such as vertical-cavity surface-emitting lasers[2] function as interfaces between electronics and optics with high data throughput (typically more than 10^9 bits/s), the optical computing technique must maintain the data flow. Although adjustment is an unavoidable task in an optical setting, it increases system cost, so a method with less sensitivity in adjustment is preferable. The method should extract the potential capabilities of the optical technologies that are utilized. In addition, as can be seen from other technological fields, concrete applications promote the progress of the field.

With the above background, the author presents a method for string alignment by use of a moiré technique[3] applied to one-dimensional spatial encoding patterns.[4] String alignment is one of the essential operations in genome information processing including DNA structure analysis[5] and studies of molecular evolution.[6] The method is expected to provide high data throughput by optoelectronic implementation with less-precise adjustment. It effectively utilizes the parallelism of optical processing. The importance of genome information processing is obvious.

String alignment is an operation by which the similarity of one one-dimensional sequence of data (called a string) to another sequence is evaluated by a matching process. Each datum is a member of a set consisting of a finite number of elements. For DNA the elements are four bases: adenine (A), guanine (G), thymine (T), and cytosine (C); for protein they are 20 amino acids.[7] Because DNA stores all the information about an organism, analysis of the DNA sequence yields important knowledge in the life sciences. During the process of evolution, DNA sequences have been modified by cross-over or mutation. Thus the various string alignments in DNA provide information on the evolution of organisms.

Figure 1 shows an example of string alignment of two strings. Because of exchange, deletion, and insertion of elements, matching must be applied locally, and the possibility of a lateral shift of the elements in the strings must be considered. In addition, the lengths of the strings to be examined are from several hundreds to several billions of elements. As a result, this operation is relatively troublesome in spite of its simple appearance. A typical solution to this problem is based on dynamic programming.[8] Although various implementations have been presented,[9] they are basically sequential, and their performance is restricted. If a new algorithm in which the two-dimensionality of optics is considered, we can implement an effective solution to the problem. Based on this idea, the author presents a method that uses an optical computing technique.

Figure 2 illustrates the processing procedure of string alignment by use of a spatial coding and moiré technique. First, the strings to be aligned are encoded into coded images. Figure 3 shows coding patterns for the bases in DNA. Each element in the string is replaced by any one of the patterns and stretched in the longitudinal direction. A collective set of the patterns constitutes a coded image. After the coding, the coded images overlap at a small intersection angle. Then bright segments will appear where the sequences of elements in both strings are identical. Therefore we can detect the positions of the matched sequences in the strings by observing the moiré pattern, and the information can be utilized for identifying the common sequence in the strings.

As features of the moiré technique, the processing results can be obtained even with rough adjustment. Changing the intersection angle between the coded images enables us to control the number of data in or the sensitivity of the operation. The larger the angle is, the more data are processed at a time with less sensitivity. With a lateral shift of one coded image, we can change the positional correspondence between the strings in a matching operation. By use of a lateral

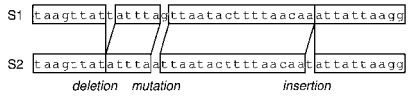

Fig. 1. String alignment of two DNA sequences.

0146-9592/99/231681-03$15.00/0

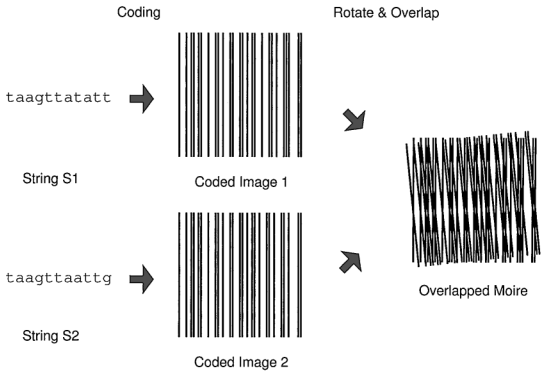

Fig. 2. Processing procedure of the proposed method.

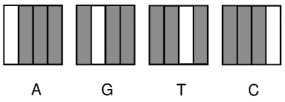

Fig. 3. Coding patterns for DNA bases.

shift, alignment even for long strings can be achieved effectively.

To verify the proposed method, string alignment between two DNA-simulated sequences was demonstrated. Figure 4 shows two strings to be aligned, which are a part of the DNA of *Mycoplasma genitalium*[10] (S1) and its modification (S2). S2 is generated by insertion of A at the 181st and the 301st loci and deletion of two bases (C and T) at the 241st and the 242nd loci of S1. Both strings are 500 elements long. Spatial coding in Fig. 3 was applied to the strings, and the coded images were printed on a transparent sheet and on an opaque sheet of paper.

Figure 5 shows the patterns observed when the coded images are overlapped with pairs of [S1, S1] and [S1 , S2]. The former corresponds to the case of exact matching, in which a single long line appears; the latter is the case of matching between locally matched strings. As one can see from the figure, four segments are found with three gaps. Note that the output pattern contains multiple matching results for different loci with different amounts of shift. The lateral position shows the locus of the string, and the longitudinal position indicates the relative amount of shift for the matching operation. Note that the operation achieved is nothing but parallel processing implemented by an optical computing technique. Figure 6 depicts an observed pattern with lateral shift of one coded image. In this case the positional correspondence between the strings can be changed.

Note that the output pattern gives only rough information on local matching between the strings. However, considering the size of data to be processed, such rough information is useful for reducing the target set for subsequent strict processing. For example, the output pattern can be captured by a two-dimensional photodetector array; then the signals of the bright segments are used for precise matching by a digital processor. In addition, we can analyze the pattern directly by visual inspection. This might be useful as a convenient test for string alignment.

The processing capability of the proposed method is evaluated by the number of data on the overlapped images. Figure 7 shows a simplified model of the overlapped strips for the coded images. Each line of the

```
taagttattatttagttaatacttttaacaatattattaaggtatttaaaaaatactatt
atagtatttaacatagttaaatacccttccttaatactgttaaattatattcaatcaatac
atatataatattattaaaatacttgataagtattatttagatattagacaaatactaatt
ttatattgctttaatacttaataaatactacttatgtattaagtaaatattactgtaata
ctaataacaatattattacaatatgctagaataatattgctagtatcaataatttactaat
atagtattaggaaaatacccataataatatttctacataatactaagttaatactatgtgt
agaataataaataatcagattaaaaaaattttatttatctgaaacatatttaatcaattg
aactgattattttcagcagtaataattacatatgtacatagtacatatgtaaaatatcat
taatttctgttatatataat
```

S1

```
taagttattatttagttaatacttttaacaatattattaaggtatttaaaaaatactatt
atagtatttaacatagttaaatacccttccttaatactgttaaattatattcaatcaatac
atatataatattattaaaatacttgataagtattatttagatattagacaaatactaatt
attatattgctttaatacttaataaatactacttatgtattaagtaaatattactgtaata
--aataacaatattattacaatatgctagaataatattgctagtatcaataatttactaat
aatagtattaggaaaatacccataataatatttctacataatactaagttaatactatgtgt
agaataataaataatcagattaaaaaaattttatttatctgaaacatatttaatcaattg
aactgattattttcagcagtaataattacatatgtacatagtacatatgtaaaatatcat
taatttctgttatatataat
```

S2

Fig. 4. String data for experimental alignment.

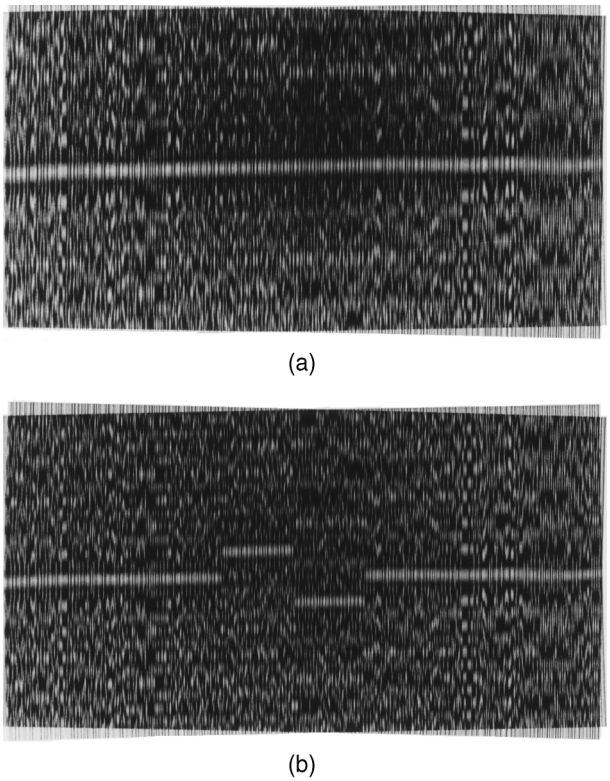

Fig. 5. Output patterns obtained by evaluation on (a) S1 and S1 and (b) S1 and S2.

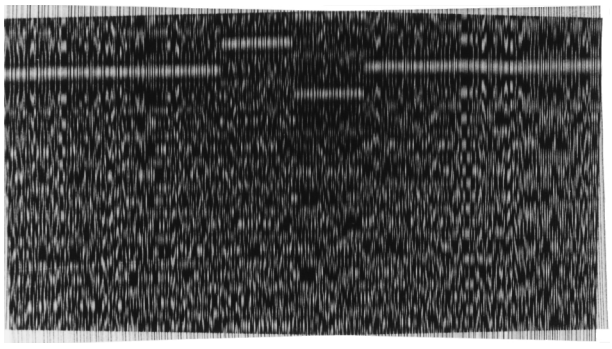

Fig. 6. Modified output pattern obtained with the lateral shift of one image.

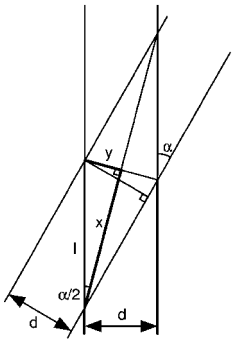

Fig. 7. Simplified model of the moiré structure.

strip pattern corresponds to a stretched coding pattern. Line spacing d is identical to that of the string elements. In this model a crossing point of the lines indicates a position where a matching operation is executed. A rhombus enclosed by crossing lines provides a unit area of one matching operation. Lengths x and y in Fig. 7 are calculated as follows:

$$x = \frac{d}{2\sin(\alpha/2)}, \tag{1}$$

$$y = \frac{d}{2\cos(\alpha/2)}, \tag{2}$$

where α is the intersection angle. Thus the unit area occupied by one matching operation S is

$$S = 2xy = d^2/\sin\alpha. \tag{3}$$

The total number of matching operations on an image plane P is approximated by

$$P = LMd^2/S = LM\sin\alpha \tag{4}$$

for the case that M elements are arranged on the image with each coding pattern stretched a length Ld.

Inasmuch as a moiré pattern is observed for $x >> d$, α must be much smaller than $\pi/3$. Assuming that the double inequality is satisfied for a $4\times$ difference, α should be smaller than $\pi/12$. Thus the maximum number of matching operations $P_{\max}$ is estimated to be $0.26LM$. Considering that M is the number of elements arranged on the image, the maximum number of matching operations for one element $Q_{\max}$ is $0.26L$. In other words, $Q_{\max}$ corresponds to the range of the area for the parallel matching operation. For example, $L = 500$ and $M = 500$ yields $P_{\max} = 65,000$ and $Q_{\max} = 130$. If we assume a vertical-cavity surface-emitting laser (VCSEL) array for data display, more than 10^{14} operations/s of processing throughput are expected.

As an implementation of the method, an optical system consisting of a one-dimensional VCSEL array, a transparent type of spatial light modulator, and a two-dimensional photodetector array is appropriate. The VCSEL and the spatial light modulator display the target strings, and the photodetector array captures the output pattern. The output signal is processed by a postdigital processor to produce exact correspondence of the strings. Although the refresh rate of the spatial light modulator is not so fast as that of the VCSEL, the nature of string alignment relaxes the speed requirement. For a string alignment, a specific short string is compared with many short strings or with a long string. In this case, if at least one string can be changed at a high data rate, the processing throughput of the system is expected to remain high.

In terms of the coding format, various implementations exist. For example, binary coding can reduce the required area for each code and increase data density. For the DNA case, only two subcells are sufficient for binary coding, whereas four subcells are used in Fig. 3. This coding format is especially useful for amino acid alignment in which 20 kinds of element should be identified. Another option is a coding rule in which the chemical characteristics of each element are considered. Polarity and electric charge of amino acids are important factors in investigation of a protein. If similar code patterns are mapped to similar properties, this method can be extended to process information on the attributes of the target strings.

In conclusion, a method for string alignment that uses a spatial coding and moiré technique has been presented. Assuming an application to DNA sequencing, the principle has been verified by experiments. With a simplified model, the processing capability was estimated. This method is expected to be a practical optical computing technique in which optical and electronic technologies are used effectively.

The author's e-mail address is tanida@mls.eng.osaka-u.ac.jp.

References

1. Optical Society of America, *Optics in Computing*, Technical Digest (Optical Society of America, Washington, D.C., 1999).
2. K. Lear, A. Mar, K. D. Choquette, S. P. Kilcoyne, R. P. Schneider, and K. M. Geib, Electron. Lett. **31,** 886 (1996).
3. O. Bryngdahl, J. Opt. Soc. Am. **64,** 1287 (1974).
4. J. Tanida and Y. Ichioka, Int. J. Opt. Comput. **1,** 113 (1990).
5. J. Meidanis, J. C. Setabal, and J. C. Setubal, *Introduction to Computational Biology* (PWS-Kent, Boston, Mass., 1996).
6. R. Lewin, *Patterns in Evolution—The New Molecular View* (Freeman, New York, 1997).
7. B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology* (Garland, New York, 1998).
8. S. B. Needleman and C. D. Wunsch, J. Mol. Biol. **48,** 443 (1970).
9. T. F. Smith and M. F. Waterman, J. Mol. Biol. **147,** 195 (1981).
10. C. M. Fraser *et al.*, Science **270,** 397 (1995).