Available online at www.scholarsresearchlibrary.com

Scholars Research Library

Der Pharmacia Lettre, 2015, 7 (7):250-255
(http://scholarsresearchlibrary.com/archive.html)

ISSN 0975-5071
USA CODEN: DPLEB4

Position matrix in DNA sequence alignment with gaps

M. Yamuna and A. Elakkiya

School of Advanced Sciences, VIT University, Vellore, India

ABSTRACT

Many questions related to drug discovery and drug designs which are structure based can be answered using structural bioinformatics tools. A search of sequence databases followed by sequence alignment and analysis can easily answer questions related to the specificity of a particular target in a given organism. We isolate human DNA sequences, translate it into amino acids, then model a human protein structure based on the known structure of modal organism, which leads to discovering drugs that bonds the model proteins. In an alignment one may achieve better correspondence between two sequences if we allow gap in one of the sequences. In DNA sequence alignment various techniques are available. A method using binary sequence will be computational friendly. In this paper we provide a method of determining local and global DNA sequence alignment using binary matrix.

Keywords: DNA, Sequence Alignment, Local Alignment, Global Alignment, Gap

INTRODUCTION

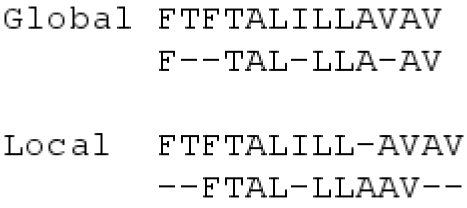
In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences [1]. A DNA sequence is presented as a sequence of characters, which may be 'A', 'G', 'C' or 'T'. To align two DNA sequences, some gaps may be inserted to sequences so that two sequences have the same length [2].

Shannon I. Steinfadt et.al presented a data-parallel algorithm for local sequence alignment based on the Smith-Waterman algorithm has been adapted for an associative model of parallel computation known as ASC. The algorithm finds the best local alignment in O (m+ n) time using m + 1 processing elements [3]. Zheng Zhang, et.al introduced a new greedy alignment algorithm with particularly good performance and showed that it computes the same alignment as does a certain dynamic programming algorithm, while executing over 10 times faster on appropriate data [4]. Wang Liang, Zhao KaiYong presents a new avenue to build more effective DNA alignment methods [5].

MATERIALS AND METHODS

Sequence Alignment

Alignment is the result of a comparison of two or more gene or protein sequences in order to determine their degree of base or amino acid similarity. Sequence alignments are used to determine the similarity, homology, function or other degree of relatedness between two or more genes or gene products. Alignments can be global or local [6]. Snapshot – 1 provides an example of global and local alignment [7].



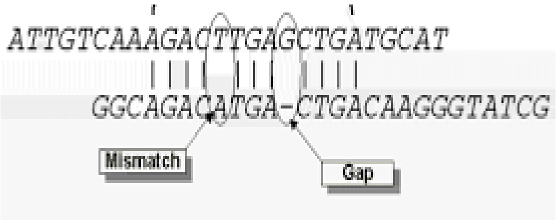
```
Global  FTFTALILLAVAV
        F--TAL-LLA-AV

Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```

Snapshot – 1

Match
Every element in a trace is either a match or a gap. Where a residue in one of two aligned sequences is identical to its counterpart in the other the corresponding amino-acid letter codes in the two sequences are vertically aligned in the trace [6].

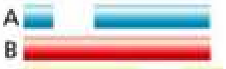
Gaps
When a residue in one sequence seems to have been deleted since the assumed divergence of the sequence from its counterpart, its "absence" is labelled by a dash in the derived sequence. When a residue appears to have been inserted to produce a longer sequence a dash appears opposite in the unaugmented sequence. Since these dashes represent "gaps" in one or other sequence, the action of inserting such spacers is known as gapping [6]. Snapshot – 2 provides an example of gap and mismatches [7].



Snapshot – 2

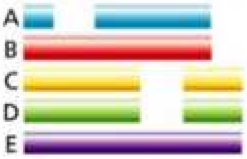
Gap Penalty
The gap penalty is a scoring system used in bioinformatics for aligning a small portion of genetic code, more accurately, fragmented genetic sequence, also termed, reads against a reference genetic sequence (e.g. The Human Genome) [8].

Pairwise Sequence Alignment
It is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid) [6]. Snapshot – 3 provides an example of pair wise sequence alignment [7].



Snapshot – 3

Multiple Sequence Alignment (MSA)
It is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied [6]. Snapshot – 4 provides an example of pair wise multiple sequence alignment [7].



Snapshot – 4

Genomic alignment
Genomic alignment tools concentrate on DNA (or to DNA) alignments while accounting for characteristics present in genomic data [6].

RESULTS AND DISCUSSION

In this section we propose a method of sequence alignment with gap, using binary matrices.

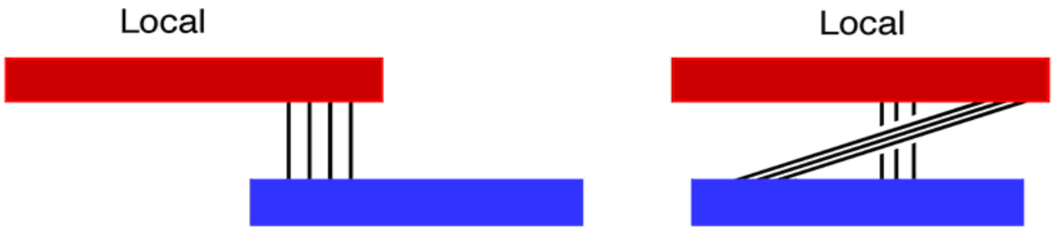
Construction of Position Matrix

Let the sequences to be aligned be S_1, S_2 of length n, m respectively, n ≤ m. We create a position matrix P of order 4 x m. The rows of the matrix represent one of A, T, G, C. Each column has exactly one non zero entry. For the j^{th} column, this nonzero entry corresponds to the j^{th} entry in the sequence. If the j^{th} column of the sequence is 1000, 0100, 0010, 00010, if the j^{th} entry in the sequence is A, T, G, C respectively. This simply means that a column with 1000 represents A, a column with 0100 represents T, a column with 0010 represents G, a column with 0001 represents C in P. For example for the sequence AATGCCT the corresponding P matrix is

$$\left[\begin{array}{ccccccc} A & A & T & G & C & C & T \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{array}\right]$$

Local Alignment

Sequences which are suspected to have similarity or even dissimilar sequences can be compared with local alignment method. It finds the local regions with high level of similarity. Snapshot –5 provides illustration of local alignment [9].



```
                  tccCAGTTATGTCAGgggacacgagcatgcagagac
                     ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Snapshot – 5

Local Alignment Using P Matrix

Let P_1 and P_2 represent the position matrices of sequences S_1 and S_2 respectively. Let the columns of P_1 be labeled as C_{11}, C_{12}, …, C_{1n} and the columns of P_2 be represented as C_{21}, C_{22},…, $C2_m$.

Step 1 Start with C_{11}. Compare C_{11}, C_{21}. If the binary sequences match, then compare C_{12} and C_{22}. If the match then continue this until the binary sequences is in the corresponding columns match.

Step 2 Let the first mismatching columns be C_{1i}, C_{2i}. This means that the binary sequences in columns C_{1i}, C_{2i} do not match. Then verify if C_{1i}, $C_{2\,i+1}$ match. If not find the first column in P_2 which matches with C_{1i}. Say Column C_{1i} and C_{2j}, i < j matches. Continue with step 1 stating with C_{1i} and C_{2j}.

After Step 2 we notice that there is a gap from C_{2i} to $C_{2\,j-1}$.

Step 3 Continue this procedure until we cannot continue any further.

In this procedure say we have reached column C_{1k}. We next will look for a column in P_2 whose binary sequence matches with the sequence in C_{1k}. If P_2 does not contain ay further column which matches this, then we can continue further with $C_{1\,k+1}$. So a gap will be created at position C_{1k}. Since only four binary sequences are possible this gap size cannot exceed 3.

For example let S_1: TACTCACGGATGATTTAGAGGCC and S_2: ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGGA be the two sequences to be compared for alignment.

Sequence - 1 and its corresponding position matrix P_1

$$\left[\begin{array}{ccccccccccccccccccccccc} T & A & C & T & C & A & C & G & G & A & T & G & A & T & T & T & A & G & A & G & G & C & C \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right]$$

Sequence - 2 and its corresponding position matrix P_2

$$\left[\begin{array}{ccccccccccccccccccccccc} A & C & T & A & C & T & A & G & A & T & T & A & C & T & T & A & C & G & G & A & T & C & A \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right.$$

$$\left.\begin{array}{ccccccccccccccccccc} G & G & T & A & C & T & T & T & A & G & A & G & G & C & T & T & G & G & A \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

As per the procedure we start the column wise matching. The first two columns of S_1 do not match with S_2. So the matching starts from column 3 of S_2. Four columns match with each other. Column 5 entry of S_1 is 0001. But column 7 entry of S_2 is 1000. The binary strings do not match. So we search for the next column in S_2 with the string entry 0001. This occurs at column 13 of S_2. We continue this procedure of gap determination until we reach column 23 of S_1. The corresponding binary string is 0001.But after gap determinations there is no more columns in S_2 with a binary value 0001. So a new gap is created (red one). Also since S_1 ends here the remaining part of S_2 is a gap as observed.

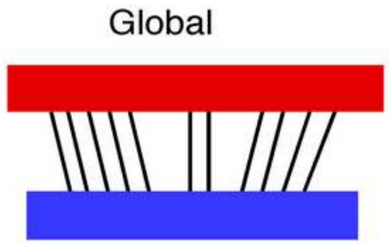
In the above matrix red and blue column represents mismatch and gap between two sequences.

Aligned Sequences

```
- - T A C T - - - - - - C - - A C G G A T - - G - - A - T T T A G A G G C C - - - - -
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
A C T A C T A G A T T A C T T A C G G A T C A G G T A C T T T A G A G G C - T T G G A
```

Global Alignment

Sequences having same length and quite similar are very much appropriate for global alignment. Here the alignment is carried out from beginning of the sequence to end of the sequences to find out the best possible alignment. Snaphshot – 6 provides illustration of local alignment [9].



```
--T--CC-C-AGT--TATGT-CAGGGGACACG—A-GCATGCAGA-GAC
  |   || |   ||   | | | |||     || | | |   | ||||    |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG—T-CAGAT--C
```

Snapshot – 6

In global matching final part of the sequences should match. So we modify the above procedure slightly for this purpose.

We first start with column C1n. Compare with column C_{2m}. Then adopt the above procedure in the reverse order ie., with $C_{1\ n-1}$, $C_{2\ m-1}$, …. until we reach a column where the strings mismatch with each other. We then continue the tracing from Step 1 ie., from C_{11}, C_{21} and continue to trace the gaps.

For example let S_1: ACTACTAGATTACGGATCGTACTTTAGAGGCTTGCACCA and S_2: ACTACTAGATTACTTACTGGATCATGTACTTTAGAGGCTGCACCA be the two sequences to be compared for global matching. We start with the last two columns. We observe that the last six columns match. We stop our process there and continue from the first columns using the usual procedure. The gaps are as seen in the matrices. Observe that when we reach column 33, there is no more matching available in S_2 as the remaining string is already matched. This mismatching is seen in red color.

Sequence - 1 and its corresponding position matrix P_1

$$\begin{bmatrix} A & C & T & A & C & T & A & G & A & T & T & A & C & G & G & A & T & C & G & T & A & C & T & T \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

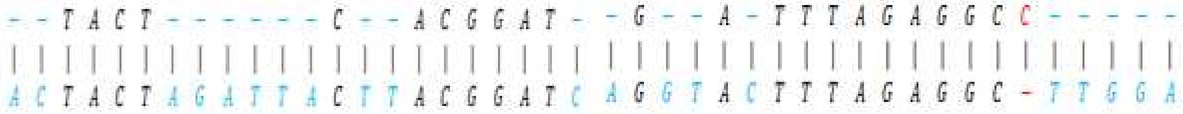
$$\begin{bmatrix} T & A & G & A & G & G & C & T & T & G & C & A & C & C & A \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Sequence – 2 and its corresponding position matrix P_2

$$\begin{bmatrix} A & C & T & A & C & T & A & G & A & T & T & A & C & T & T & A & C & T & G & G & A & T & C & A \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} T & G & T & A & C & T & T & T & A & G & A & G & G & C & T & G & C & A & C & C & A \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Aligned Sequences:



```
- - T A C T - - - - - - C - - A C G G A T - - G - - A - T T T A G A G G C C - - - - -
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
A C T A C T A G A T T A C T T A C G G A T C A G G T A C T T T A G A G G C - T T G G A
```

CONCLUSION

Protein and DNA sequences of different organisms are often related and they indicate the knowledge about species. The similarity or differences found would help in categorizing the common characteristics of species and their behaviors. In bioinformatics analysis, the identification of gaps is important in sequence matching as they could represent a mutation that could offer insight to the discovery of new functionality. The effectiveness and efficiency of gap pattern discovery relies on the development of novel pattern redundancy concepts to generate a large number of redundant flexible gap patterns. This has an application in developing drugs and treatments for diseases that have unknown regulatory functions. Minimizing the gap penalty is an issue. The proposed method guarantees that the gap penalty cannot exceed the value obtained in this method. It uses binary matrices and hence computation friendly.

REFERENCES

[1] https://en.wikipedia.org/wiki/Sequence_alignment.
[2] http://www.ebi.ac.uk/Tools/psa/.
[3] www.cs.kent.edu/~ssteinfa/papers/CASB_final.pdf.
[4] Zheng zhang, Scott Schwartz, Lukas Wagner, and Webb Miller, *Journal Of Computational Biology*, **2000,** Volume 7, Pp. 203–214
[5] arxiv.org/ pdf /1307.0194.
[6] http://www.ebi.ac.uk/Tools/psa/.
[7] https://www.google.co.in/?gws_rd=ssl#q=pairwise+sequence+alignment+definition.
[8] https://en.wikipedia.org/wiki/Gap_penalty.
[9] https://www.pitt.edu/~mcs2/teaching/biocomp/tutorials/global.html.