Available online at www.scholarsresearchlibrary.com

Scholars Research Library

Der Pharmacia Lettre, 2015, 7 (12):392-398
(http://scholarsresearchlibrary.com/archive.html)

Scholars Research Library
ISSN 0975-5071
USA CODEN: DPLEB4

DNA gap penalty using directed graphs

M. Yamuna

SAS, VIT University, Vellore

ABSTRACT

For two sequences in the alignment that share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels. Very short or very similar sequences can be aligned by hand; however, most interesting problems require the alignment of lengthy, highly variable or extremely numerous sequences that cannot be aligned solely by human effort. Instead, human knowledge is primarily applied in constructing algorithms to produce high-quality sequence alignments. A variety of computational algorithms have been applied to the sequence alignment problem Graph theory is developing as a promising field in various applications. In this method a method of penalty determination using graph theory is proposed, which can be developed into an algorithm.

Key words: DNA, Gap Penalty.

INTRODUCTION

DNA graph penalty determination is a problem studied by various researchers. Different methods and algorithms are devised for this purpose. In [1] a Fast Optimal Global Sequence Alignment Algorithm, FOGSAA, which aligns a pair of nucleotide/protein sequences faster than any optimal global alignment method including the widely used Needleman-Wunsch (NW) algorithm is proposed. Details regarding sequence alignments and various techniques is provided in [2]. In [3] a method of gap penalty using matrices is proposed. Use of graph theory in such cases is not in wide use. In this paper we propose a method of determining constant gap penalty using graph theory.

Preliminary Note

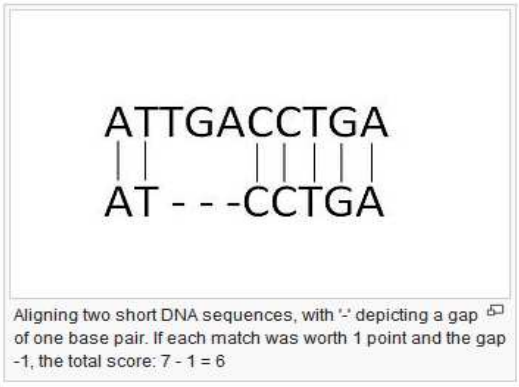
In this section some basic results required for the proposed method is provided.

DNA graph 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). The biological process of protein synthesis namely, transcription and translation or DNA replication can produce errors resulting in mutations in the final nucleic acid sequence. Therefore, in order to make more accurate decisions in aligning reads, mutations are annotated as gaps in the sequence. Gaps are penalized via various Gap Penalty scoring methods. Gaps in a DNA sequence result from either insertions or deletions in the sequence, sometimes referred to as indels.

Constant

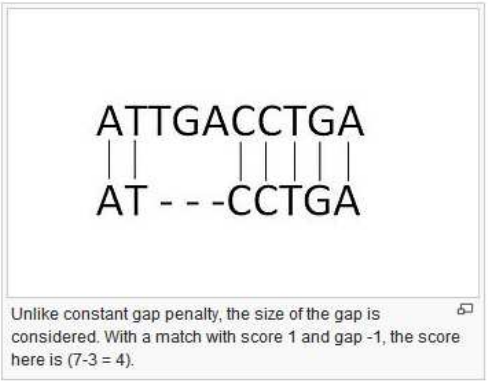
This is the simplest type of gap penalty, a fixed negative score is given to every gap, regardless of it length.



Aligning two short DNA sequences, with '-' depicting a gap of one base pair. If each match was worth 1 point and the gap -1, the total score: 7 - 1 = 6

Linear

Compared to the constant gap penalty, the linear gap penalty takes into account the length (L) of each insertion/deletion in the gap. Therefore, if the penalty for each inserted/deleted element is B and the length of the gap L; the total gap penalty would be the product of the two BL. This method favors shorter gaps, with total score decreasing with each additional gap [4].



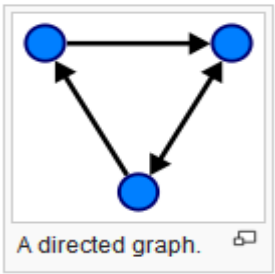
Unlike constant gap penalty, the size of the gap is considered. With a match with score 1 and gap -1, the score here is (7-3 = 4).

Digraph

In mathematics, and more specifically in graph theory, a directed graph (or digraph) is a graph, or set of vertices connected by edges, where the edges have a direction associated with them. In formal terms, a directed graph is an ordered pair G = (V, A with

- V a set whose elements are called vertices, nodes, or points;
- A set of ordered pairs of vertices, called arrows, directed edges (sometimes simply edges with the corresponding set named E instead of A), directed arcs, or directed lines.

It differs from an ordinary or undirected graph, in that the latter is defined in terms of unordered pairs of vertices, which are usually called edges, arcs, or lines [5].



A directed graph.

Snapshot – 1

Complete Symmetric Digraph

A graph G is said to be complete if there is a directed edge between every pair of vertices. Snapshot – 1 is an example of a complete directed graph with 3 vertices. A complete symmetric digraph if there is an between every

pair of vertices. Also for every edge directed between a and b, there is an edge directed between b and a. A complete symmetric digraph with four vertices is seen in Fig. 1 [6]

Graph Determination of Constant DNA Gap Penalty
Construction of Base Digraph
We consider directed symmetric digraph with 4 vertices and 16 directed graphs. Label the vertices of the graph as A, T, G, C. A sample DNA penalty graph is seen in Fig. 1

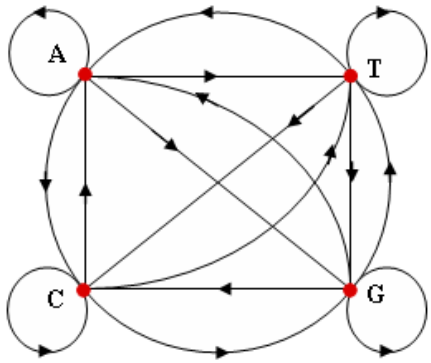


Fig. 1

We shall use this graph as the base graph for determining constant gap penalty.

Determination of Constant Gap Penalty
Construction of Digraph for a Given Sequence
Consider any random DNA sequence of any length.

For example let us consider the sequence as AATGCGCATGCA.

As in the usual determination of constant gap penalty split this into bits of size 2.

AA AT TG GC CG GC CA AT TG GC CA. If the sequence is of length k, then we generate k – 1 pairs.
For each of these pairs assign values as 1, 2,…, k – 1.

AA	AT	TG	GC	CG	GC	CA	AT	TG	GC	CA
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
1	2	3	4	5	6	7	8	9	10	11

Draw directed edge between these pairs and assign labels as 1, 2, …, k – 1. To make things comfortable we stick on to the base graph in Fig. 1and just label the edges. So if an edge has weight assigned we consider it as an edge. Edges without weights are dummy edges. For the sequence considered the resulting graph is as seen in Fig. 2

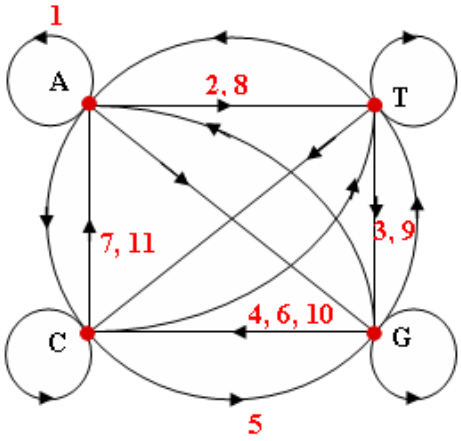


Fig. 2

In Fig. 2 AA AT TG GC CG CA are original edges. The remaining edges are dummy edges.

Gap Penalty Determination Using Graph
We consider two random sequences and construct the corresponding digraphs.

For example let us consider another sequence of same length as TATGAACATAAT. Split this into segments of length 2 to generate TA AT TG GA AA AC CA AT TA AA AT. Then assign number values for this graph sequence to generate

TA	AT	TG	GA	AA	AC	CA	AT	TA	AA	AT
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
1	2	3	4	5	6	7	8	9	10	11

The resulting graph for this sequence is seen in Fig. 3.

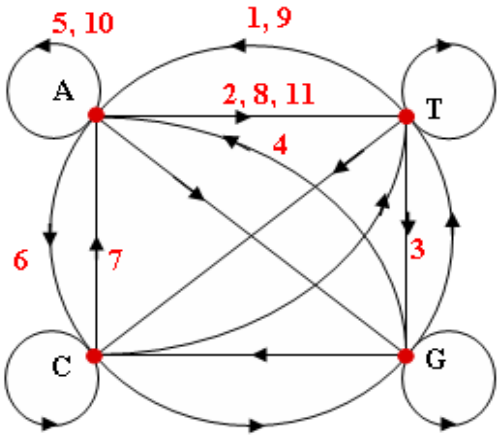


Fig. 3

When there is a gap between in some sequence while drawing the graph care is taken so that the original sequence length is maintained. For example if the two sequences are

TGCTGGAATGCA
AGCT - - AAAACA
Then we split them into two bit sequences as follows

TG	GG	GC	CT	TG	GG	GA	AA	AT	TG	GC
AG	GC	CT	-	-	-	AA	AA	AA	AC	CA

We observe that if the sequence is of length k, then the number of pairs possible is k – 1. So when there is a gap care is taken that this count is maintained so that the resulting weights of the graphs can be compared. Assigning values to these pairs the resulting sequence is

TG	GG	GC	CT	TG	GG	GA	AA	AT	TG	GC
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
AG	GC	CT	-	-	-	AA	AA	AA	AC	CA
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
1	2	3	4	5	6	7	8	9	10	11

While assigning weights to the second graph we skip weights 4, 5, 6. The graph for the second sequence is seen in Fig. 4. Observe that there are no edges with weights 4, 5, 6 in the graph.

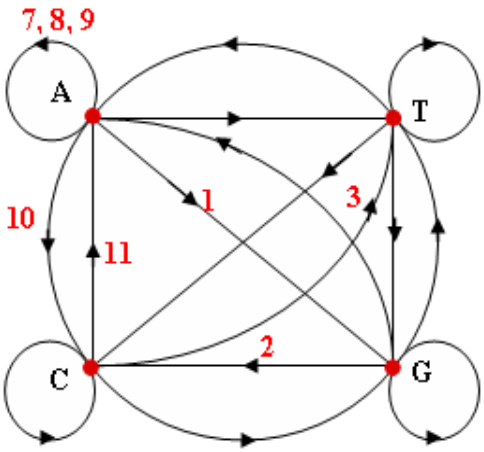


Fig. 4

Compare the graphs of the resulting sequences. We look at the edge weights of the corresponding edges and list out those edges for which the values are the same. The sum of these values is the number of positions in which the sequence match with each other. In all the remaining places they do not match. This means that the number of weights at which they do not match gives the gap penalty value.

For our two sequences we shall compare the edge values.

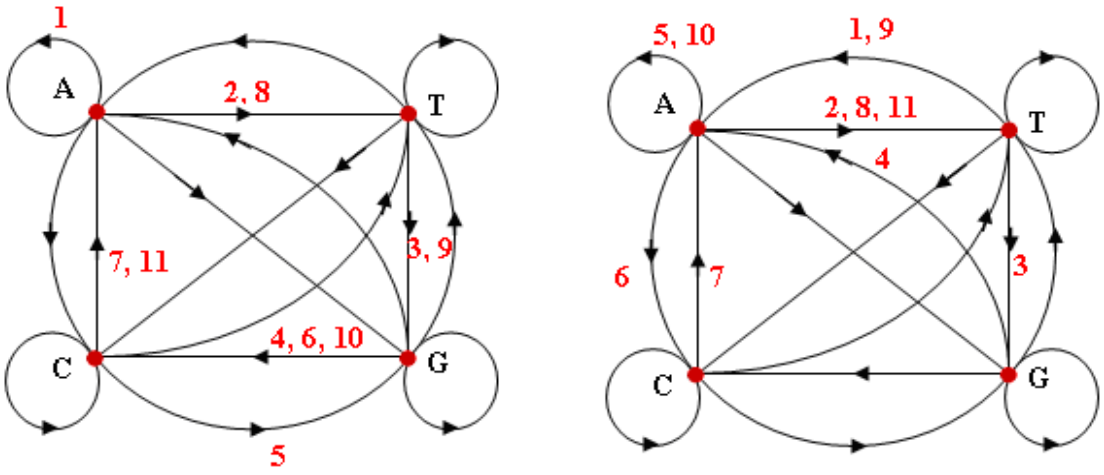


Fig. 5

Looking at both the graphs we observe that the matching edge values are 2, 3, 7, 8. In all the remaining positions they do not match. This can be seen to be true by comparing the two sequences together.

AA	AT	TG	GC	CG	GC	CA	AT	TG	GC	CA
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
1	2	3	4	5	6	7	8	9	10	11
TA	AT	TG	GA	AA	AC	CA	AT	TA	AA	AT
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
1	2	3	4	5	6	7	8	9	10	11

If the edge value is 2, this means that there is an edge drawn from the vertex preceding it to the vertex following it. So it means that in positions 2, 3 the sequences match each other. Based on this for out given sequences they match in position pairs (2, 3), (3, 4), (7, 8), (8, 9), ie., the sequences match at positions 2, 3, 4, 7, 8, 9. This can be verified by comparing the sequences.

A	A	T	G	C	G	C	A	T	G	C	A
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
T	A	T	G	A	A	C	A	T	A	A	T
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
1	2	3	4	5	6	7	8	9	10	11	12

So the gap penalty is 12 – 6 = 6.

Summarizing we determine the gap penalty as follows

Step 1 Consider the two sequences S1, S2. Let the length of the longest sequence be p.
Step 2 Draw the weighted digraphs G1, G2.
Step 3 Determine the positions at which same edges have same weights. List the weights in increasing order say w1, w2, …, wk (These are numbers).
Step 4 Determine the list (w1, w2), (w2, w3), …, (wk, wk+1).
Step 5 Create the list of these weights in distinct increasing order. This is the matching list. Let the length of this list be t.
Step 6 Gap Penalty = p – t.
Illustration

Table – 1 provides some sample sequences and their gap penalties

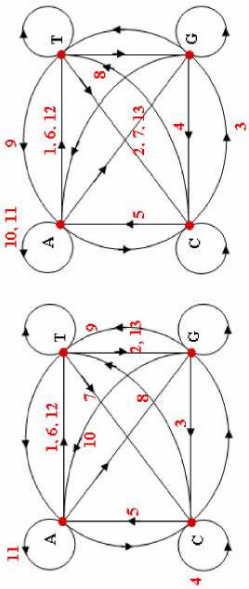
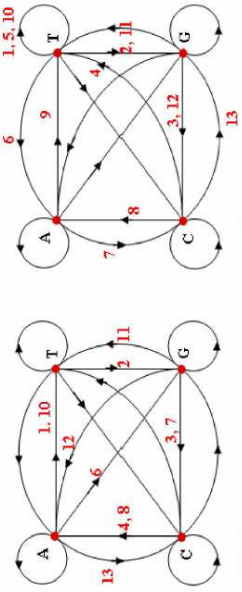
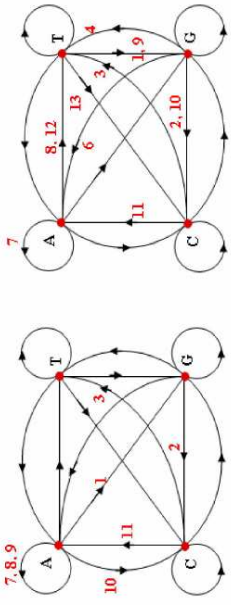
S. No	Sequences	digraphs	Matching weight List	Gap Penalty
1	ATGCCATCTGAATG ATCGCATCTAAATC		1,5,6,7,11,12	5
2	ATGCAAGCAATGAC TTGCTTACATTGCG		2,3,8,11	7
3	TGCTGGAATGCATC AGCT - - AAAACA - -		2,3,7,11	7

Table – 1

CONCLUSION AND FURTHER WORK

The given sequences may be of any length, but the graphs under consideration is only with 4 vertices and 16 edges. So even if the sequence is long, we need to compare only 16 edges for common weights in them. This makes the proposed method simple. So the proposed method can be used for determining gap penalty. This method can be developed as an algorithm and hence used for gap penalty determination.

REFERENCES

[1] Angana Chakaraboty, Sanghamitra Bandyopadhayay, FOGSAA: Fast Optimal Global Sequence Alignment Algorithm, Scientific Reports 3, **2013**.
[2] http://www.dnabaser.com/articles/sequence%20alignment/
[3] M. Yamuna, A. Elakkiya, Position Matrix in DNA Sequence Alignment with Gaps, Der Pharmacia Lettre, **2015**, 7 (7) 250 – 255.
[4] https://en.wikipedia.org/wiki/Gap_penalty
[5] https://en.wikipedia.org/wiki/Directed_graph
[6] Narasing Deo, Graph theory with application to Engineering and Computer Science, Prentice Hall India (**2010**).