

Genomic Sequence Analysis Using Gap Sequences and Pattern Filtering

Shih-Chieh Su, Chia H. Yeh and C. -C. Jay Kuo

Department of Electrical Engineering, University of Southern California, CA, USA
shihchis@usc.edu, {chyeh,cckuo}@sipi.usc.edu

Abstract

A new pattern filtering technique is developed to analyze the genomic sequence in this research based on gap sequences, in which the distance of the same symbol is recorded consecutively as a sequence of integers. Sequence alignment and similarity testing can be performed on a family of gap sequences over selected patterns. The gap sequence offers a new way for sequence structural analysis. The match between the gap sequences is considered as a frame match while a true match requires both frame and stuffing match. Simulation results show that the extension of gap match indicates the corresponding segment extension in the original genomic sequence. Thus, we are able to generalize the conventional alignment and scoring methods in a more adaptive way.

1. Introduction

Genomic sequences are composed of symbols. For example, there are four deoxyribonucleic acid bases, *i.e.* adenine (a), cytosine (c), guanine (g), and thymine (t), to construct the DNA sequences. Furthermore, the protein can be analyzed into twenty amino acid building blocks. Therefore, both DNA and protein sequences can be represented by an alphabet containing a finite number of characters.

The set of symbols does not define an algebraic structure in the sense that there is no meaningful symbol-to-symbol operation other than the simple comparison operation. It is difficult to define sequence properties such as the spectrum and the variance for such symbolic sequences. If we can translate the genomic sequence of symbols into numbers in a meaningful way, it will be easier to derive rich features in the number domain.

There are several ways to translate symbols into numbers such as the direct symbol mapping [1] and the direct vector mapping [2]. However, the translated number serves only as an index, to which commonly used arithmetic operations cannot be properly applied.

In this research, we extract the structural information from genomic sequences using the gap sequence. Each number in the gap sequence stands for the gap length between two successive occurrences of the same selected

pattern. New similarity measurements can be defined for the gap sequence, and used for the matching and alignment purpose. Simulation results are provided to demonstrate that the gap sequence can be a tool for genomic DNA sequence analysis.

2. Pattern filtering and gap sequences

A structure mapping technique, called pattern filtering, is introduced to keep only the structural information in the translated sequence. We start out with a simple case of pattern filtering. Let S be a DNA sequence of length n and $S[i]$ the DNA character at location i of sequence S . Suppose that we select the single DNA character 'a' to be the pattern that has the length $j=1$. The indicator sequence $I_a[i]$ is a binary sequence that $I_a[i]=1$ for a pattern hit at location i , and $I_a[i]=0$ otherwise. To recover the pattern locations in $S[i]$ from $I_a[i]$, two pseudo hits are added to both ends of $I_a[i]$, *i.e.* $I_a[0]=I_a[n-j+2]=1$. Finally the 'a'-filtered gap sequence $F_a[i]$ is generated by cumulating zeros between two ones in $I_a[i]$. The relationship between above sequences is demonstrated in Fig. 1.

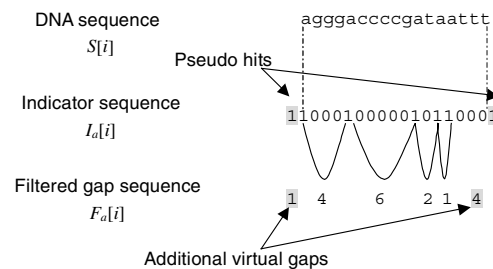


Fig. 1: The pattern filtering process

3. Post processing

Each value in the gap sequence indicates the interval between two occurrences of the selected pattern. In most cases, the Poisson distribution is suitable to model the intervals of occurrences. By examining histograms of gap sequences for several genomic sequences as shown in Fig. 2, we find that larger gap values have a stronger discriminating power and can be used to locate possible matches.

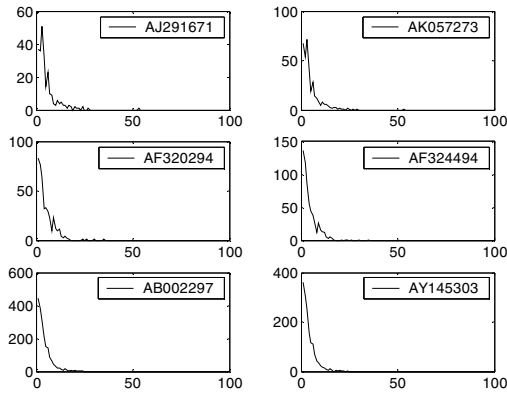


Fig. 2: Histograms of 'a'-filtered gap sequences

The proposed histogram-aided pairwise sequence alignment (HAPA) takes the following as the input: query sequence S_q , target sequence S_t , the associating F_q, F_t , with their histograms H_q, H_t ; the minimum pivot p_{min} ; the minimum matching length m_{min} . We do the following:

- 1) Select pivot p as the maximum common gap value in H_q and H_t . If $p < p_{min}$, then abort.
- 2) Try to extend the matching location in the unmasked parts of F_q and F_t . If the matching length is greater than m_{min} , the match is qualified and we mask the matching segments in F_q and F_t . Go to Step 1.

We can further generalize this algorithm to a multiple sequence alignment version under the same concept. While the systems in the BLAST family have a difficult time in finding the short pattern match in the database, the proposed method examines only the pivot. Another generalization can be made upon the qualifying factor for a matching. Assuming that there are totally K matching segments collected during the processing of the HAPA algorithm. We define the column vectors $\mathbf{b}_k = [p_k, l_k]^T$ and $\mathbf{w}_k = [w_{pk}, w_{lk}]^T$. The value p_k denotes the value of pivot that leads to the k^{th} match, while l_k being the matching length of the match. The coefficients w_{pk} and w_{lk} are used to weight p_k and l_k , respectively. The matching score of the k^{th} matching segment is $S_w(k) = \mathbf{w}_k^T \mathbf{b}_k$.

4. Simulation results

We test the histogram-aided sequence alignment algorithm using the set of DNA sequences {AB018272(A), BC022783(B), AK023845(C)} downloadable from the NCBI GenBank. The result with parameters $p_{min}=10$, $m_{min}=5$ for this set is demonstrated in Fig. 3. A circle is located at the top of the pivot for each match. More than alignment, the result also gives the structural information about the selected pattern in these DNA sequences.

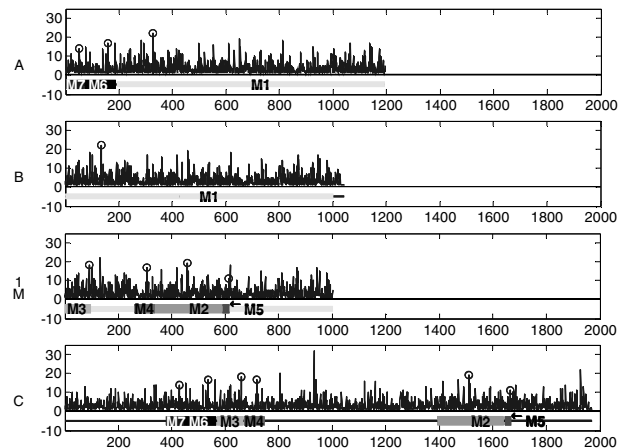


Fig. 3: The histogram-aided alignment result

In order to verify the correctness to frame matching using HAPA and gap sequences, we conducted the following test over a selection of DNA sequences. The program first determined all segments satisfying parameters p and m_{min} , and then check the frame-matched segments for exact match. The result is summarized in Table 1.

p	m_{min}	Frame matched pairs	Exact matched pairs
5	5	1608	874(54%)
	10	365	363(99%)
	20	8	8(100%)
10	5	139	129(93%)
	10	128	118(92%)
	20	25	25(100%)

Table 1: The results of the gap match test

5. Conclusion

In this research, we proposed a frame matching technique for genomic sequence analysis using gap sequences and the HAPA algorithm. We studied the behavior of these gap sequences. Simulation results demonstrated that the performance of HAPA can find similar segments very fast. One major contribution of this research is that, given partial knowledge of a genomic sequence segment, we can still predict the remaining portions of this segment accurately.

6. References

- [1] W. Wong and D. H. Johnson, "Computing linear transforms of symbolic signals", *IEEE Transaction on Signal processing*, vol. 50, 2002, pp. 628-634.
- [2] T. Kahveci and A. K. Singh, "An efficient index structure for string databases", *VLDB*, 2001, pp. 351-360.