BLAST: From Dropoff Score to Dropoff Percentage

Deepak Garg[1], S C Saxena [2] and L M Bhardwaj[3]

[1,2]Thapar Institute of Engg. And Technology, Patiala

[3]Biomolecular Electronics & Nanotechnology CSIO, Chandigarh

Abstract

Basic Local Alignment Search Tool Rapidly identifies statistically identical patterns between known nucleotide, protein or amino acid sequences. As the speed in which sequences are increasing is very fast, the sequence analysis tools have to be sensitive to this fact, so as to remain in use. BLAST is one of the widely used sequence analysis tools. In this paper we are proposing an improvement in one of the parameters of BLAST. Currently BLAST is using drop-off score to calculate the highest scoring pairs between two sequences. A change has been proposed to calculate the threshold score that determines the inclusion of the subsequence in the result. Instead of using a drop-off score, if we use a drop-off percentage, it gives better results for some sequences.

Keywords

BLAST, Drop-off score, Sequence Alignment, Algorithm Design

1. Introduction

BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) is a set of similarity search programs designed to explore all of the available (DNA and protein) sequence databases. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. BLAST uses the concept of a "segment pair" which is a pair of sub-sequences of the same length that form an ungapped alignment. The algorithm first looks for short words that are present in both sequences and then extends these at either end to find the longest segments present in both. The statistical significance of these High-scoring Segment Pairs is evaluated to determine whether the matches are random or not. Thus, the scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background.

1.1 X, drop-off

It is the value of score, which tells how much the score is allowed to drop off since the last maximum. If X value is set high the quality of the alignment is degraded, on the other hand if smaller value is set for X, there are chances of missing some alignment.

The drawback in this approach is that the value of X depends on the substitution scores, gap initiation and extension costs. So the easier way to calculate the drop off will be if we can define some drop off percentage. Drop off percentage will be the number of mismatches allowed after some significant number of matches. In this case there will be no need to refer to substitution matrix and hence there will be increase in the speed.

To make the concept clearer we will try to align two sentences. To keep the example simple we will ignore spaces and wont allow gaps in the alignment. Here are the two sentences:

THE QUICK BROWN FOX JUMPS OVER THE LAY DOG.

THE QUIET BROWN CAT PURRS WHEN SHE SEES HIM.

Here, the two sentences first have six matches

THE QUI
THE QUI

And then two mismatch before the next match.

CK B
ET B

For this, the drop off percentage comes out to be (2 / 6) *100 = 33.33 %. So if the X value is kept around this value, it will solve our purpose. Lets assume it to be 35%. So next time in this alignment, if the drop off percentage comes out to be greater than 35%, the extension will be terminated and the alignment will be trimmed back to the last match. Next we have five matches before a mismatch

BROWN F
BROWN C

And then we have continuously four mismatches before a match

FOX JU
CAT PU

So the drop off percentage comes out to be (4 / 5) * 100 = 80%. This is greater than 35%. So the extension is terminated at this point and is trimmed back to N, the last match. The whole process is shown graphically in the Figure 1.

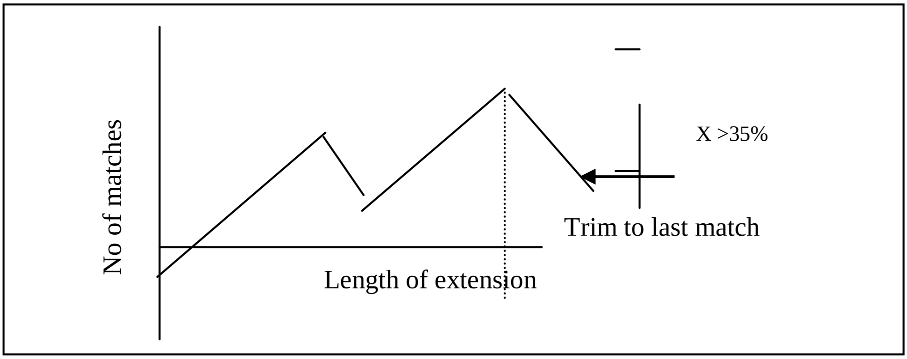

Figure 1. Drop-off percentage vs. drop-off score

The results that come out with drop off percentage are same as with drop off. The drop off also terminates the extension at N but the approach behind termination is different. To explain the approach we will try to align the same two sentences using a scoring scheme in which identical letters score +1 and mismatches score –1. To keep the example simple we will ignore spaces and wont allow gaps in the alignment. Although only extension to the right is shown, it also occurs to the left of the seed. Here a variable X that represents drop off score must be selected. It represents how much the score is allowed to drop off since the last maximum. Let's set X to 5 and see what happens. Here we have to keep track of the sum score and drop off score.

```
THE QUICK BROWN FOX JUMP
THE QUIET BROWN CAT PURR
123   45 654  5 6 7 8 9  876  5 6 5 4   << score
000   00 012 1 0 0 0 0  123   4 3 4 5   << drop off score
```

The maximum score for this alignment is 9, and the extension is terminated when the score drops to 4. After terminating the alignment is trimmed back to the maximum score. The maximum score was at N, so the alignment ends at N. The whole process is shown graphically in the following Figure 2.

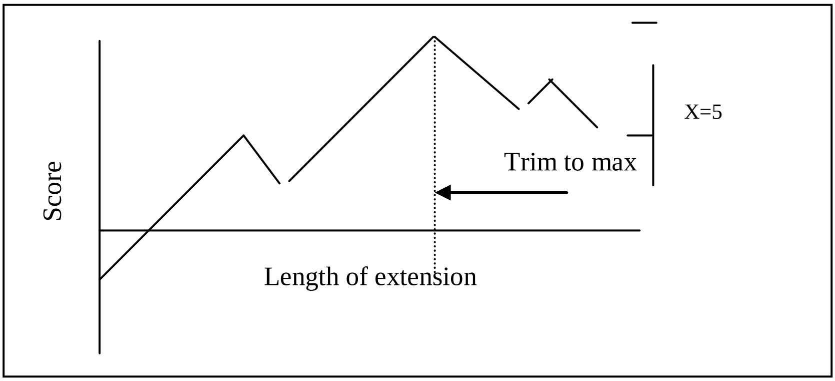

Figure 2

The alignment produced here is same as produced by our proposed approach. It also aligns the two sentences up to N. But it involves the sum score also which regularly needs to refer to scoring matrix and hence slows down the speed of the search.

2. Conclusion

The drop off parameter's value depends on the substitution scores, gap initiation and extension costs. It regularly needs to refer to scoring matrix and hence it slows down the speed of the search. So the easier way to calculate the drop off is suggested. We redefined the drop off to drop off percentage. Drop off percentage is the number of mismatches allowed after some significant number of matches. In this case there is no need to refer to substitution matrix and hence there is increase in the speed. The results that come out with drop off percentage are same as with drop off but the approach behind drop off percentage is different.

References

[1] Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994)*Nature Genet*., **6**, 119–129.

[2] Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F.(1998) *Nucleic Acids Res*., **26**, 1–7.

[3] Bergerat,A., de Massy,B., Gadelle,D., Varoutas,P.C., Nicolas,A. and Forterre,P. (1997) *Nature*, **386**, 414–417.

[4] Bevan,M., Hilbert,H., Braun,M., Holzer,E., Brandt,A., Duesterhoeft,A.,Hoheisel,J., Jesse,T., Heijnen,L., Vos,P., *et al*. (1998) GenBank accession no. 2961386.

[5] Bevan,M., Hilbert,H., Braun,M., Holzer,E., Brandt,A., Duesterhoeft,A.,Hoheisel,J., Jesse,T., Heijnen,L., Vos,P., *et al*. (1998) GenBank accessionno. 2961387.

[6] Buehler,E., Dewar,K., Feng,J., Kim,C., Li,Y., Shinn,P., Sun,H., Conway,A.,Conway,A., Kurtz,D., *et al*. (1997) GenBank accession no. 2213598.

[7] Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G.,Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., *et al*. (1996)*Science*, **273**, 1058–1073.

[8] Chinnaiyan,A.M., Chaudhary,D., O'Rourke,K., Koonin,E.V. andDixit,V.M. (1997) *Nature*, **388**, 728–729.

[9] Collins,J.F., Coulson,A.F.W. and Lyall,A. (1988) *Comp. Appl. Biosci*., **4**,67–71.

[10] Gumbel,E.J. (1958) *Statistics of Extremes*. Columbia University Press,New York, NY.

[11] Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**,10915–10919.

[12] Klenk,H.P., Clayton,R.A., Tomb,J., White,O., Nelson,K.E., Ketchum,K.A.,Dodson,R.J., Gwinn,M., Hickey,E.K., Peterson,J.D., *et al*. (1997) *Nature*,**390**, 364–370.

[13] Koonin,E.V., Mushegian,A.R., Galperin,M.Y. and Walker,D.R. (1997)*Mol. Microbiol*., **25**, 619–637.

[14] LeBlanc,D.J., Lee,L.N. and Inamine,J.M. (1991) *Antimicrob. AgentsChemother*., **35**, 1804–1810.

[15] Li,P., Nijhawan,D., Budihardjo,I., Srinivasula,S.M., Ahmad,M.,Alnemri,E.S. and Wang,X. (1997) *Cell*, **91**, 479–489.

[16] Mott,R. (1992) *Bull. Math. Biol*., **54**, 59–75.

[17] Mushegian,A.R., Bassett,D.E., Jr, Boguski,M.S., Bork,P. and Koonin,E.V. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 5831–5836.

[18] Nagase,T., Seki,N., Tanaka,A., Ishikawa,K. and Nomura,N. (1995)*DNA Res*., **2**, 167–174.

[19] Pearson,W.R. (1998) *J. Mol. Biol*., **276**, 71–84.

[20] Robinson,A.B. and Robinson,L.R. (1991) *Proc. Natl Acad. Sci. USA*, **88**,8880–8884.

[21] Seshagiri,S. and Miller,L.K. (1997) *Curr. Biol*., **7**, 455–460.

[22] Smith,T.F., Waterman,M.S. and Burks,C. (1985) *Nucleic Acids Res*., **13**,645–656.

[23] Staden,R. (1989) *Comp. Appl. Biosci*., **5**, 89–96.

[24] Tsui,H.T., Mandavilli,B.S. and Winkler,M.E. (1992) *Nucleic Acids Res*., **20**, 2379.

[25] Waterman,M.S. and Vingron,M. (1994) *Stat. Sci*., **9**, 367–381.

[26] Wilson,R., Ainscough,R., Anderson,K., Baynes,C., Berks,M., Bonfield,J.,Burton,J., Connell,M.,Copsey,T., Cooper,J., *et al*. (1994) *Nature*, **368**,32–38.

[27] Wootton,J.C. and Federhen,S. (1993) *Comp. Chem*., **17**, 149–163.

[28] Yue,D., Maizels,N. and Weiner,A.M. (1996) *RNA*, **2**, 895–908.44 Dracheva,S., Koonin,E.V. and Crute,J. (1995) *J. Biol. Chem*., **270**,14148–14153.

[29] Zhang,Z., Berman,P. and Miller,W. (1998) *J. Comput. Biol*., **5**, 197–210.

[30] Zou,H., Henzel,W.J., Liu,X., Lutschg,A. and Wang,X. (1997) *Cell*, **90**,405–413.