FramePlus - a sensitive algorithm for aligning DNA to protein sequences

Eran Halperin, Simchon Faigler and Raveh Gill-More
Compugen Ltd., 72 Pinchas Rosen st., Tel-Aviv, Israel 69512.
Corresponding author: raveh@compugen.co.il

Biological sequence alignment algorithms have become extremely popular in the last few years, and are now being used by thousands of researchers as arguably the most important annotation tool in bioinformatics. However, the search algorithm used may have a crucial effect on the success of an annotation project: different algorithms will find (and miss) different hits under different circumstances.

Frame algorithms are a special case of sequence alignment algorithms, when one wishes to compare a nucleic acid sequence to an amino acid sequence. They are particularly useful for annotating Expressed Sequence Tags (ESTs).

The first frame algorithm developed was *Translated Search* [1] (also known as *six-frame translation*), and is directly based on the Smith-Waterman algorithm [2]. Heuristic database search packages have incorporated frame algorithms early on [3, 4].

However, early frame algorithms were not very error tolerant, especially when the errors cause frameshifts. A solution to this was introduced by *FrameSearch* [5]. In this work we introduce a new frame algorithm, termed *FramePlus*, which is an extension of *FrameSearch* where we model sequencing errors separately from indels of amino-acids caused by evolution. Since these are two different phenomena it is reasonable to expect that this better modeling will result in increased sensitivity.

In order to test this conjecture, we have utilized and customized ideas by Brenner et al. [6] for algorithm benchmarking based on the *SCOP* database [7, 8] of structurally classified proteins, and implemented a general framework for benchmarking frame algorithms. We used this framework to compare all the above mentioned algorithms.

Our results suggest that *FramePlus* is significantly more sensitive than other algorithms, and in cases of low sequence identity may find as much as 14% more true hits than any of the other tested algorithms. Although *FramePlus* is slower than heuristic algorithms such as *BlastX* when implemented on a standard computer, it can be accelerated by up to 3 orders of magnitude on special purpose hardware. The *FramePlus* source code are freely available, at `ftp.compugen.co.il/pub`

References

[1] Peltola, H., Soderlund, H., and Ukkonen, E. 1986. Algorithms for the search of amino acid patterns in nucleic acid sequences. *Nucl. Acid. Res* **14**:99–107.

[2] Smith, T., and Waterman, M. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147**:195–197.

[3] Pearson, W., and Lipman, D. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**:2444–2448.

[4] Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**:403–410.

[5] Edelman, I., Faigler, S., Mintz, E., Natan, A., and Devereux, J.R. 1995. A rigorous alignment program for searching protein databases with nucleic acid queries. Poster, Genome Sequence and analysis Conference, Hilton-Head, Florida.

[6] Brenner, S.E., Chothia, C., and Hubbard, T.J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.* **95**:6073–6078.

[7] Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**:536–540.

[8] `http://scop.mrc-lmb.cam.ac.uk/scop`