Bioinformatics, 32(2), 2016, 304–305
doi: 10.1093/bioinformatics/btv575
Advance Access Publication Date: 1 October 2015
Applications Note

Sequence analysis

ALP & FALP: C++ libraries for pairwise local alignment *E*-values

Sergey Sheetlin[1], Yonil Park[1], Martin C. Frith[2,†] and John L. Spouge[1,*,†]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, MD 20894, USA and [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Koto-ku, Tokyo 135-0064, Japan

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.
Associate Editor: John Hancock

Received on August 21, 2015; revised on September 25, 2015; accepted on September 28, 2015

Abstract

Motivation: Pairwise local alignment is an indispensable tool for molecular biologists. In real time (i.e. in about 1 s), ALP (Ascending Ladder Program) calculates the *E*-values for protein–protein or DNA–DNA local alignments of random sequences, for arbitrary substitution score matrix, gap costs and letter abundances; and FALP (Frameshift Ascending Ladder Program) performs a similar task, although more slowly, for frameshifting DNA-protein alignments.
Availability and implementation: To permit other C++ programmers to implement the computational efficiencies in ALP and FALP directly within their own programs, C++ source codes are available in the public domain at http://go.usa.gov/3GTSW under 'ALP' and 'FALP', along with the standalone programs ALP and FALP.
Contact: spouge@nih.gov
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Finding and aligning similar sequences is arguably the most fundamental task in bioinformatics. The main method for analysing biosequences (novel proteins, ancient DNA, environmental metagenomics DNA, single-molecule sequencing data, whole genomes, etc.) is to compare them to sequence databases or to each other. It is important to know whether a similarity is significant, i.e. unlikely to occur by chance alone. For gapless alignments, *E*-value estimation is rapid because of analytic formulas (Karlin and Altschul, 1990), but for gapped alignment it has traditionally required hours to days to pre-calculate the statistical parameters for arbitrary scoring schemes and letter abundances (Altschul *et al.* 2001). Many alignment programs therefore limit users to the specific substitution score matrices, gap costs and letter abundances corresponding to the pre-calculations, with adverse consequence to the sensitivity of sequence searches (Eddy, 2009), particularly in applications [e.g. (Kuznetsov, 2011)] where the limitations are inappropriate.

Supplementary Table S1 compares BLAST and other alignment programs (Buchfink *et al.*, 2015; Edgar, 2010; Harris, 2007; Hauswedell *et al.*, 2014; Noe and Kucherov, 2005; Somervuo and Holm, 2015; Suzuki *et al.*, 2015; Zhao *et al.*, 2012). BLAST does not allow scoring schemes usefully tailored to searching for remote protein homology (Yamada and Tomii, 2014), remote DNA homology (Chiaromonte *et al.*, 2002), AT-rich genomes (Frith, 2011), or bisulphite-converted DNA (Frith *et al.*, 2012). In particular, because of skewed compositions (Bastien *et al.*, 2005; Paila *et al.*, 2008), BLAST can yield misleading p-values relevant to the genomes of clinically important pathogens like AT-rich malaria or GC-rich tuberculosis. Several other alignment programs also allow only a limited set of scoring schemes.

2 Methods and features

The ALP and FALP software libraries can calculate *E*-values for any scoring scheme and letter abundances (Park *et al.*, 2009; Sheetlin

et al., 2014). ALP received its name because it uses ascending ladder scores in a random sum to compute its statistical parameters. The limiting distribution of random local alignment scores approximates a Gumbel distribution when a scoring matrix, gap penalty, and letter abundances are in the 'logarithmic phase' (Arratia and Waterman, 1985). In a rapid preliminary computation, ALP diagnoses whether the input parameters (the scoring scheme and letter abundances) are in or effectively too close to the logarithmic phase for computational purposes. If the computation is feasible, ALP then computes *E*-values in real time (i.e. in about 1 s) for most logarithmic scoring schemes, although as the input parameters approach the boundary of the logarithmic phase, its computational time lengthens. In any case, ALP simulates to estimate modified Gumbel parameters and their standard errors from the formulas in the Supplementary Materials, permitting computations of alignment *E*- and *P*-values.

For shorter sequences, alignment scores deviate increasingly from a Gumbel distribution, hence several 'finite size' corrections (FSCs) have been proposed. ALP incorporates a recent FSC (Park *et al.*, 2012).

ALP allows insertion costs to differ from deletion costs. This has useful applications. For example, PacBio and nanopore DNA sequences have different insertion and deletion error rates, so accurate alignment should use different insertion and deletion costs. ALP also allows asymmetric score matrices, and different letter abundances in the two sequences. This is useful for comparing proteomes or genomes with differing compositions (Bastien *et al.*, 2005; Paila *et al.*, 2008).

The FALP library provides similar functionality for DNA-protein alignments with frameshifts. It allows arbitrary genetic codes, and does not assume that translated DNA has typical amino acid abundances. FALP's pre-calculations are slower than ALP's (typically, minutes instead of fractions of seconds). In practice, however, FALP is much faster than the computations required when aligning modern multi-gigabase datasets.

3 Conclusions

The ALP and FALP programs give researchers the ability to compute reliable statistical significances for local alignment without limiting them to particular scoring schemes or sequence compositions; the ALP and FALP libraries permit programmers to develop C++ local alignment tools that can compute p-values themselves. These libraries have already been incorporated in the LAST aligner (http://last.cbrc.jp) (Kielbasa *et al.*, 2011).

Funding

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and National Center for Biotechnology Information (S.S., Y.P.,and J.L.S.) and by KAKENHI Grant Number 26700030 (M.C.F.).

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.*, **29**, 351–361.

Arratia,R. and Waterman,M.S. (1985) Critical phenomena in sequence matching. *Ann. Prob.*, **13**, 1236–1249.

Bastien,O. *et al.* (2005) Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions. *C R Biol.*, **328**, 445–453.

Buchfink,B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods.*, **12**, 59–60.

Chiaromonte,F. *et al.* (2002) Scoring pairwise genomic sequence alignments, *Pac. Symp. Biocomput.*, 115–126.

Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Frith,M.C. (2011) A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res.*, **39**, e23.

Frith,M.C. *et al.* (2012) A mostly traditional approach improves alignment of bisulfite-converted DNA. *Nucleic Acids Res.*, **40**, e100.

Harris,R.S. (2007) Improved pairwise alignment of genomic DNA. Phd Thesis. Center for Comparative Genomics and Bioinformatics. The Pennsylvania State University.

Hauswedell,H. *et al.* (2014) Lambda: the local aligner for massive biological data. *Bioinformatics*, **30**, i349–i355.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, **87**, 2264–2268.

Kielbasa,S.M. *et al.* (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.*, **21**, 487–493.

Kuznetsov,I.B. (2011) Protein sequence alignment with family-specific amino acid similarity matrices. *BMC Res. Notes.*, **4**, 296

Noe,L. and Kucherov,G. (2005) YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.*, **33**, W540–W543.

Paila,U. *et al.* (2008) Genome bias influences amino acid choices: analysis of amino acid substitution and re-compilation of substitution matrices exclusive to an AT-biased genome. *Nucleic Acids Res.*, **36**, 6664–6675.

Park,Y. *et al.* (2012) New finite-size correction for local alignment score distributions. *BMC Res. Notes*, **5**, 286–286.

Park,Y. *et al.* (2009) Estimating the gumbel scale parameter for local alignment of random sequences by importance sampling with stopping Times. *Ann. Stat.*, **37**, 3697–3714.

Sheetlin,S.L. *et al.* (2014) Frameshift alignment: statistics and post-genomic applications. *Bioinformatics*, **30**, 3575–3582.

Somervuo,P. and Holm,L. (2015) SANSparallel: interactive homology search against Uniprot. *Nucleic Acids Res.*, **43**, W24–W29.

Suzuki,S. *et al.* (2015) Faster sequence homology searches by clustering subsequences. *Bioinformatics*, **31**, 1183–1190.

Yamada,K. and Tomii,K. (2014) Revisiting amino acid substitution matrices for identifying distantly related proteins. *Bioinformatics*, **30**, 317–325.

Zhao,Y. *et al.* (2012) RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, **28**, 125–126.