BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 19 2010, pages 2460–2461
doi:10.1093/bioinformatics/btq461

Sequence analysis

Advance Access publication August 12, 2010

Search and clustering orders of magnitude faster than BLAST

Robert C. Edgar
Tiburon, CA 94920, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Biological sequence data is accumulating rapidly, motivating the development of improved high-throughput methods for sequence classification.

Results: UBLAST and USEARCH are new algorithms enabling sensitive local and global search of large sequence databases at exceptionally high speeds. They are often orders of magnitude faster than BLAST in practical applications, though sensitivity to distant protein relationships is lower. UCLUST is a new clustering method that exploits USEARCH to assign sequences to clusters. UCLUST offers several advantages over the widely used program CD-HIT, including higher speed, lower memory use, improved sensitivity, clustering at lower identities and classification of much larger datasets.

Availability: Binaries are available at no charge for non-commercial use at http://www.drive5.com/usearch

Contact: robert@drive5.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 27, 2010; revised on August 3, 2010; accepted on August 6, 2010

1 INTRODUCTION

Biological sequence data is accumulating more quickly than the growth in computing efficiency predicted by Moore's Law (Butte, 2001), which motivates the development of improved high-throughput methods. Computational sequence analysis often begins by classifying sequences by reference to a database or by *de novo* classification (clustering), which is used to predict homology and function, reduce redundancy, generate subsets that are tractable for more computationally expensive methods, compare data from different environments and quantify ecosystem diversity. Central to most such classification methods is a sequence search algorithm such as BLAST (Altschul *et al*., 1990), which has been widely adopted for its high speed and sensitivity.

2 METHODS

UBLAST and USEARCH are new algorithms for sequence database search that seek high-scoring local and global alignments, respectively. UCLUST is a new clustering algorithm that employs USEARCH as a subroutine to assign sequences to clusters. High-throughput is achieved by using a fast heuristic designed to enable rapid identification of one or a few good hits rather than all homologous sequences. For a given query, database sequences are sorted in order of decreasing number of words in common to exploit the fact that similar sequences tend to have short words in common (see e.g. Edgar, 2004a). When examined in this order (i) if a hit exists in the database, it is likely to be found among the first few candidates, and (ii) the probability that a hit exists falls rapidly as the number of failed attempts increases. A search can therefore often be terminated after examining a small number of candidates without a large cost in sensitivity. Pair-wise sequence comparisons are performed using standard fast alignment techniques, including gapless high-scoring segment pair detection by extending word seeds and banded dynamic programming. More details are provided in the Supplementary Material.

Table 1. Comparison on Pfam and Rfam

DB	Method	Seqs./ sec.	Sens (Err) all	Sens (Err) Med.	Sens (Err) Low
Pfam	ublast	55	97 (2)	96 (1)	75 (5)
	usearch	77	92 (3)	70 (3)	22 (8)
	blastp	0.16	99 (2)	96 (3)	98 (3)
Rfam	ublast	684	99.4 (0.6)	–	96 (4)
	usearch	671	99.3 (0.7)	–	93 (7)
	blastn	2.3	99.4 (0.6)	–	97 (3)
	mega-blast.	8.8	73 (0.2)	–	35 (0)

Seqs./sec is the throughput, not including time required to load the database into memory. Sensitivity (Sens.) and error rate (Err.) are expressed as percentages and defined by comparing the families of the sequence pair in the top hit: if these are in the same (different) family, the hit is considered to be a true (false) positive. The medium-identity subset ('Med') is defined as 40–50% for Pfam.; low-identity ('Low') is 20–35% for Pfam and <75% for Rfam.

3 RESULTS

I compared UBLAST and USEARCH to NCBI BLAST+ v2.2.22 using Pfam-A v24.0 (Finn *et al*., 2008) and Rfam v9.1 (Gardner *et al*., 2009), two large sequence family databases containing 4.9M protein domains and 192k RNAs, respectively. In each case, 1000 sequences were extracted at random to use as a query and the remainder used as a search database (Table 1 and Supplementary Table S1). As a simple model of a classification task, only the top hit was considered, which was designated as a true (false) positive if it belonged to the same (different) family as the query, though sequences in different families may in fact be homologous. On this test, UBLAST was ∼70× faster than MEGABLAST with significantly higher sensitivity, ∼300× faster than BLASTN with similar sensitivity, and ∼350× faster than BLASTP with comparable sensitivity above the protein twilight zone (20–35% id.).

To illustrate typical improvements achieved by UCLUST compared to the widely used program CD-HIT (Li and Godzik, 2006), the most efficient previous clustering method, a set of 1.1×10^6 pyrosequencing reads of length ∼300 nt was taken from a recent microbial ecology study (Costello *et al*., 2009) and clustered at representative identities (Table 2). UCLUST produced higher quality clusters, enabled clustering at lower identities, used

Table 2. UCLUST and CD-HIT compared

Set	Prog.	Id. (%)	Size	Sim. (%)	Time	Mem. (Mb)
A	uclust	75	2852	87.1	2 min 27 s	34
	cd-hit	75	(cd-hit cannot cluster at <80% id.)			
	uclust	85	536	91.5	3 min 8 s	36
	cd-hit	85	343	88.9	9 h 10 min	349
	uclust	90	230	96.0	1 min 48 s	40
	cd-hit	90	175	92.1	1 h 2 min	349
	uclust	95	73	97.7	2 min 14 s	55
	cd-hit	95	68	95.9	1 h 1 min	349
	uclust	99	11	99.5	13 min 9 s	165
	cd-hit	99	15	99.1	2 h 3 min	411
B	uclust	85	506	91.4	5 h 5 min	40
		90	223	95.9	3 h 10 min	49
		95	61	95.7	4 h 4 min	81

Set A is 1.1×10^6 reads from Costello *et al.* (2009), set B contains 100 copies of A. CD-HIT failed on set B. Prog., method; Id., clustering threshold; Size, average cluster size (bigger is considered better); Sim., average identity between a cluster member and its representative sequence (higher is considered better); Time, CPU time; Mem., maximum amount of RAM used by the program. Experiments were performed on a commodity laptop computer with 2 Gb RAM. Default options were used for both programs except to specify the identity threshold and the amount of available memory. Cluster size is divided by 100 for set B to allow direct comparison with A. The size and similarity are comparable, showing that cluster quality is only slightly degraded when processing the huge number of sequences in set B.

substantially less memory for large sets and was often one or more orders of magnitude faster. CD-HIT generated clusters with larger, and thus apparently better, average size at 99% identity, but this proved to be an artifact of a bug in CD-HIT as almost half (47%) of the reported identities were 98% and therefore fell below the threshold by the CD-HIT's own measure. Many of these assignments were verified as being below 99% by creating independent alignments using MUSCLE (Edgar, 2004b). Despite generating smaller clusters, which should enable higher average similarity to be achieved, CD-HIT produced clusters with lower average similarity than UCLUST in all cases. This shows that when a sequence can be assigned to more than one cluster, UCLUST tends to identify a better match.

To demonstrate the scalability of UCLUST, 100 copies of the Costello *et al.* (2009) set were concatenated, giving an input file with 1.1×10^8 sequences. Clustering a set of this size with previous methods would require large-scale computational resources, while UCLUST was able to generate high-quality clusters at identities between 85% and 95% in 3–5 h on a commodity laptop computer in <100 Mb memory (Table 2).

4 DISCUSSION

UBLAST and USEARCH introduce a new search paradigm that seeks one or a few good hits rather than all hits in order to improve throughput. This approach is well suited to next-generation sequence classification, where search is often a critical bottleneck in analysis. Low-ranking hits are rarely needed and may even be undesirable due to increased overhead. The user can tune the trade-off between speed and sensitivity by adjusting parameters that control how many candidate hits are examined before the search is terminated. Speed and sensitivity are thus strongly dependent on the data and program options. In typical applications, USEARCH typically achieves good sensitivity at around 40% identity and above for amino acids and 65% for nucleotides. UBLAST is sensitive to more distant relationships, with useful sensitivity extending into the twilight zone for proteins. BLASTP has significantly better sensitivity at lower identities, which is partly due to refinements of the BLAST algorithm not yet implemented in UBLAST and also the problem that some distantly related proteins have no perfectly conserved words of the required length. UBLAST and USEARCH require memory that is roughly 10× the size of the database, which is sometimes more than BLAST but in most cases is readily accommodated by currently available commodity computer hardware.

UCLUST is definitively superior to CD-HIT. It is usually significantly faster, uses significantly less memory, can cluster at lower identities and is more sensitive. While CD-HIT often fails to identify the closest cluster, or overlooks that a match is possible (false negative), UCLUST rarely misses a match and in most cases finds the best possible match. UCLUST also enables rapid clustering of much larger numbers of sequences.

UBLAST, USEARCH and UCLUST can dramatically reduce the resources required for classification of large sequence sets, and will therefore be of value to biologists in a wide range of applications.

ACKNOWLEDGEMENTS

The author is grateful to Rob Knight and Richa Agarwala for helpful discussions.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Butte,A.J. (2001) Challenges in bioinformatics: infrastructure, models and analytics. *Trends Biotechnol.*, **19**, 159–160.

Costello,E.K. *et al*. (2009) Bacterial community variation in human body habitats across space and time. *Science*, **326**, 1694–1697.

Edgar,R.C. (2004a) Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Res.*, **32**, 380–385.

Edgar,R.C. (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Finn,R.D. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.

Gardner,P.P. *et al*. (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.