BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 12 2012, pages 1645–1646
doi:10.1093/bioinformatics/bts197

Sequence analysis

Advance Access publication April 23, 2012

Tachyon search speeds up retrieval of similar sequences by several orders of magnitude

Joshua Tan[1,2], Durga Kuchibhatla[1], Fernanda L. Sirota[1], Westley A. Sherman[1], Tobias Gattermayer[1], Chia Yee Kwoh[1], Frank Eisenhaber[1,3,4], Georg Schneider[1] and Sebastian Maurer-Stroh[1,5,*]

[1]Bioinformatics Institute (BII), Agency for Science Technology and Research (A*STAR), 30 Biopolis Street, #07-01, Matrix, 138671, [2]Institute of High Performance Computing (IHPC), Agency for Science Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, 138632, [3]Department of Biological Sciences (DBS), National University of Singapore (NUS), 8 Medical Drive 4, 117597, [4]School of Computer Engineering (SCE), Nanyang Technological University (NTU), 50 Nanyang Drive, 637553 and [5]School of Biological Sciences (SBS), Nanyang Technological University (NTU), 60 Nanyang Drive, 637551, Singapore

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: The usage of current sequence search tools becomes increasingly slower as databases of protein sequences continue to grow exponentially. Tachyon, a new algorithm that identifies closely related protein sequences ~200 times faster than standard BLAST, circumvents this limitation with a reduced database and oligopeptide matching heuristic.

Availability and implementation: The tool is publicly accessible as a webserver at http://tachyon.bii.a-star.edu.sg and can also be accessed programmatically through SOAP.

Contact: sebastianms@bii.a-star.edu.sg

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

Received on October 17, 2011; revised on March 27, 2012; accepted on April 16, 2012

1 ALGORITHM AND WEB SERVER

The continuous exponential growth of sequence databases creates new challenges for bioinformatics tools where a simple BLAST search (Altschul *et al.*, 1997) of one protein against NCBI's NR (Sayers *et al.*, 2011) taking minutes can become too long to be practical. Careful study of the features and natural sampling of sequence space allows further optimization of peptide-indexing based look-up methods as oligopeptides reaching a database-specific critical length can become characteristic for a protein and its family of related sequences (see Supplementary Material).

We have developed Tachyon, a peptide-indexing based look-up method that identifies sequences similar to a user-defined query protein at unprecedented speeds and look-up times of a few hundred milliseconds, rather than minutes on a single dual-core processor. Fast peptide-indexing is not new but is in fact the underlying basis of commonly used methods such as BLAST (Altschul *et al.*, 1997), BLAT (Kent, 2002), etc. However, Tachyon achieves a substantial speed advantage by reducing the searched sequence space as a result of associating each entry in the database (e.g. NR) only with representative oligopeptides which are defined as the most frequent peptides of a defined length within the indexed database, excluding low-complexity regions [SEG 12 2.2 2.5 settings (Wootton and Federhen, 1996)]. Through extensive empirical testing of different parameters including peptide length, we have found that representing database entries by using five pentapeptides was a good compromise between search space reduction (speed) and correct sequence retrieval (sensitivity) for the large NCBI NR database. Database hits sharing a specified number of pentapeptides (default: 3 out of the 5 indexed peptides) with the query are then subjected to a more detailed search over the full length sequences to evaluate the significance of each hit. For this second step, Tachyon uses an internal default algorithm based on a quick pairwise L-mer overlap score (e.g. for $L = 5$, the number of pentapeptides shared between query and subject sequences over the total number of pentapeptides present in both sequences). The user has the additional option of running classical methods such as BLAT (Kent, 2002), FASTA and SSEARCH (Pearson, 2000) instead. The basic algorithm steps are illustrated in Figure 1.

The fastest implementation of this algorithm is achieved on hardware where the indexed database can be kept in the working memory (e.g. 2 GB for the reduced index and 20 GB for the full length sequences considering the current NR). To relieve users from special hardware requirements and the burden of constantly updating and indexing the huge NCBI NR database, we make Tachyon searches against NR available as a web service (database updated monthly) at the URL: http://tachyon.bii.a-star.edu.sg. Users can paste single sequences or small sets as well as upload FASTA-formatted files. The web interface shows hit sequences with associated scores/E-values and includes an optional hit alignment using MAFFT (Katoh *et al.*, 2005) with subsequent display in Jalview (Waterhouse *et al.*, 2009). Additional links to the FASTA sequence of hits and other databases such as GenBank, BLink, NCBI Taxonomy, PDB UniProt, as well as to ANNIE, an integrated sequence annotation suite (Ooi *et al.*, 2009), are provided. Different filters, such as limiting the display of the results to specific source databases (PDB, RefSeq and SwissProt/UniProtKB), to a number of sequences or by using keywords are available on-line. Tachyon is also accessible programmatically through SOAP and permanently connected to a compute cluster, which, together with a length-based

*To whom correspondence should be addressed.

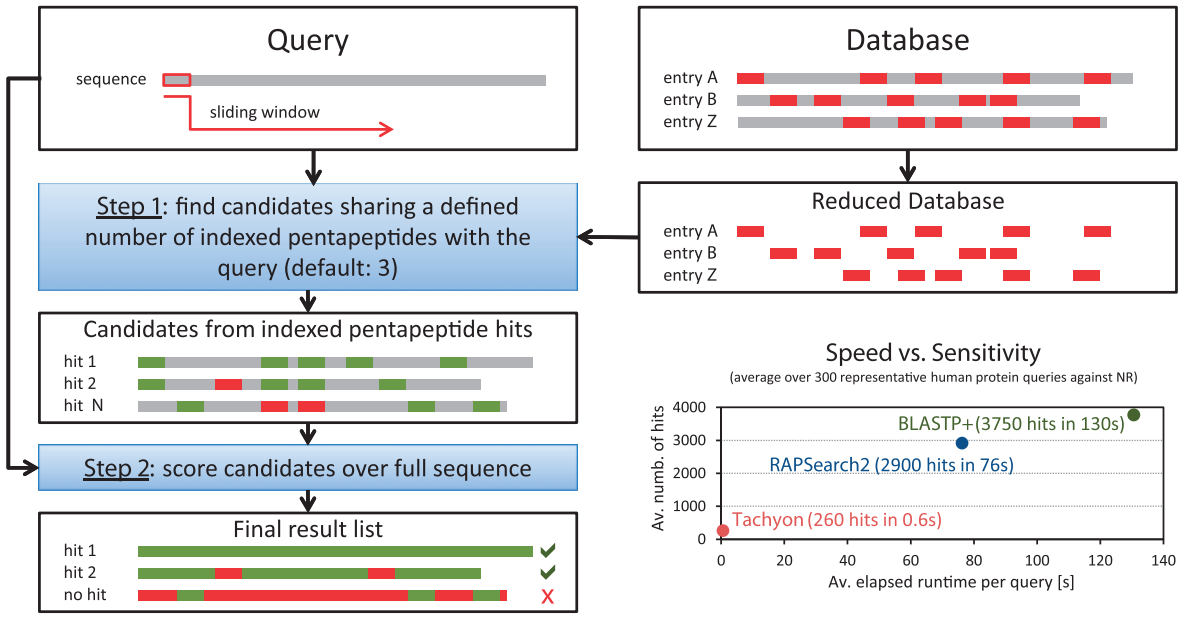

Fig. 1. Algorithm scheme and speed versus sensitivity benchmark. The searched sequence space is reduced by associating each entry in the database with only five representative pentapeptides (red rectangles, top right side). Database hits sharing at least a defined number of pentapeptides with the query (green rectangles in Step 1, left side) are then subjected to a more detailed search over the full length sequences to evaluate the significance of each hit (Step 2, bottom left side) for which multiple classical methods to choose from are implemented in the Tachyon Web Service. A benchmark for assessing the speed versus sensitivity profile of key methods (summary of results in plot in bottom right side) is described in detail in the Supplementary Material

load-balancing algorithm makes it possible to match large protein sets at high speeds. For example, searching the whole human proteome including isoforms (IPI, v3.78 and 86 702 entries) against NR on 30 Dual-Core AMD Opteron 2220 processors took only 13 min. Results obtained using SOAP can be either zipped FASTA sequences of the hits or, fastest, only the accession numbers of the best hits, which allows quick cross-connection of a query sequence to popular databases and other SOAP services.

2 BENCHMARK AND DISCUSSION

We benchmarked Tachyon against key methods with searches of 500 representative human proteins against NR and Figure 1 shows the clearly unique application spectrum arising from the speed versus sensitivity profile (detailed benchmark information in Supplementary Material). For example, at default settings, Tachyon is ~200× faster than BLASTP+ but finds ~14× less hits when searching NR. Similarly, it is ~120× faster than the recently developed fast method RAPSearch2 (Zhao *et al.*, 2012) while finding ~11× less hits. In absolute numbers, each Tachyon search took on average ~0.6 s and returned ~260 hits with the last significant hit having on average 74% identity to the query.

The considerable gain of speed despite the loss of sensitivity has many practical applications where BLAST searches would take too long or the usual sensitivity of BLAST may not be needed. These scenarios include commonly needed tasks like accession-independent linking up of a sequence to its most closely related entries in well curated databases such as GenBank (Benson *et al.*, 2011), UniProt [(The Universal Protein Resource (UniProt) in 2010, 2010] or diverse pathway and interaction databases. Our approach of fast sequence matching can also be used to rapidly connect any sequence to precalculated BLAST search results such as in BLink (Sayers *et al.*, 2011). Therefore, to allow sequence similarity searches with Tachyon-speed and BLAST-sensitivity, users can simply click in one step from the best Tachyon-hit against NR to the respective BLink entry with the corresponding full BLAST result. Additional links are available to GenBank, PDB, UniProt and to the ANNIE sequence annotation web tool.

In summary, we developed Tachyon to very quickly connect a query protein sequence to highly similar sequences and selected associated resources. It is available as a SOAP service and on-line at http://tachyon.bii.a-star.edu.sg.

Funding: This work was supported by the Agency for Science Technology and Research (A*STAR), Singapore.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Benson,D.A. *et al.* (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.

Katoh,K. *et al.* (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Ooi,H.S. *et al.* (2009) ANNIE: integrated de novo protein sequence annotation. *Nucleic Acids Res.*, **37**, W435–W440.

Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.

Sayers,E.W. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.

The Universal Protein Resource (UniProt) in 2010. (2010) *Nucleic Acids Res.*, **38**, D142–D148.

Waterhouse,A.M. *et al.* (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Meth. Enzymol*, **266**, 554–571.

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