RAFTS3: Rapid Alignment-Free Tool for Sequence Similarity Search

Ricardo Assunção Vialle[1*], Fábio de Oliveira Pedrosa[2], Vinicius Almir Weiss[1], Dieval Guizelini[1], Juliana Helena Tibaes[1], Jeroniza Nunes Marchaukoski[1], Emanuel Maltempi de Souza[2], Roberto Tadeu Raittz[1]

[1]Laboratory of Bioinformatics, Sector of Professional and Technological Education, Federal University of Paraná, Curitiba, Paraná, Brazil.
[2]Department of Biochemistry and Molecular Biology, Federal University of Paraná, Curitiba, Paraná, Brazil.

E-mails: Ricardo Assunção Vialle[1*] - ricardovialle@gmail.com

Fábio de Oliveira Pedrosa[2] - fpedrosa@ufpr.br

Vinicius Almir Weiss[1] - viniciusweiss@gmail.com

Dieval Guizelini[1] - dievalg@gmail.com

Juliana Helena Tibaes[1] - tibaes.juliana@gmail.com

Jeroniza Nunes Marchaukoski[1] - jeroniza@gmail.com

Emanuel Maltempi de Souza[2] - souzaem@ufpr.br

Roberto Tadeu Raittz[1] - raittz@gmail.com

Corresponding author: Ricardo Assunção Vialle[1*], ricardovialle@gmail.com

Keywords: alignment-free; sequence comparison; protein sequence comparison; genome annotation.

ABSTRACT (words counted: 159 / limit: 350)

Background: Similarity search of a given protein sequence against a database is an essential task in genome analysis. Sequence alignment is the most used method to perform such analysis. Although this approach is efficient, the time required to perform searches against large databases is always a challenge. Alignment-free techniques can offer alternatives for comparing sequences without the need of alignment.

Results: Here we present RAFTS3, a fast protein similarity search tool that uses a candidate selection step based on shared k-mers and a comparison measure using a binary co-occurrence matrix of amino acid residues. RAFTS3 performed searches many times faster than those with BLASTp against large protein databases, such as NR, Pfam or UniRef, with a small loss of sensitivity depending on the similarity degree of the sequences.

Conclusions: RAFTS3 offers a new alternative for fast comparison of protein sequences, genome annotation and biological data mining. The source code and the standalone files for Windows and Linux platform are available at: https://sourceforge.net/projects/rafts3/

BACKGROUND

Biological data mining deals with the discovery of patterns, trends, answers, or other meaningful information that is hidden in the data. Sequence comparison is the main component in the retrieval system from genomic databases. An efficient sequence comparison algorithm is critical for searching biological databases. Usually, bioinformatics workflows use algorithms based on sequence alignment such as BLAST [1] to search for similarity of DNA/RNA or protein sequences against large sequence databases. Comparisons involving large databases such as NCBI NR [2], however, are computationally costly and demand long running times. The development of new computationally faster algorithms may provide significant improvement in biological pattern search. A class of techniques that can speed up sequence comparison is the alignment-free approach [3].

Algorithms based on sequence alignment are efficient in detecting similarities between protein sequences. These approaches have been improved since the first methods. Originally alignment techniques used dynamic programming to produce an optimized alignment between the sequences. Although efficient implementations have been developed, the computational load to compare large amounts of sequences makes these algorithms very slow and demanding [3,4]. To compensate for the high computational cost of full alignments, heuristic approaches were proposed. In general, these methods use subsequences of pre-determined length "k" (k-mers). The subject database is searched to find sequences that have common k-mers related to the query sequence. The k-mers are then extended using scores schemes to maximize the aligned regions. However, although heuristic methods are somewhat efficient to perform searches in large databases, they also have their limitations, such as loss of sensitivity and parameter thresholds [4].

The alignment-free methods offer a way to obtain a similarity measure between sequences without the need to perform alignments. These methods are also based on the assumption that two similar sequences share a certain portion of k-mers. Given a query sequence, the alignment-free methods generally work by selecting subject

sequences with k-mers that are present in both query and subject sequences. The procedure then applies a statistical method to establish a similarity ranking for these sequences [5].

Generally, alignment-free techniques are divided in two classes: a) methods based on words (sequences) with fixed sizes, followed by the use of statistical analysis including procedures based on defined metrics such as Euclidean distance and entropy of frequency distributions; and b) methods where words of fixed sizes are not required for statistical analysis, using data compression and/or Kolmogorov complexity scale independent representations by iterated maps. Reviews of these techniques are available at [3,5,6].

Several alignment-free techniques have been proposed with different degrees of success. The very first proposal of an alignment-free method for biological sequence comparison showed to be superior to alignment based algorithms in some aspects such as the ability to compare low similarity sequences [7]. Since then, it has been applied in phylogenetic reconstruction [8–11], identification of homologous proteins [4], genome annotation [12], classification of metagenomic sequences [13], and identification of regulatory sequences [14]. Also, it has been shown as an efficient technique for sequence filtering [15].

Alignment-free approaches have been used to replace alignment based approaches for searching and comparing sequences against large databases showing significant increase in speed. PAUDA [16] is an alternative to BLASTx for searching sequencing reads against protein databases in metagenomics. PVC (Periodicity Count Value) is a method for finding homologous nucleotide sequences as alternative for BLASTn [17]. USEARCH [18] is an alternative to BLASTp that applies a k-mer approach to perform searches of protein sequences against a protein database.

In this paper we propose a fast and efficient alignment-free method named RAFTS3. The method is based on amino acid co-occurrence matrices and on a new heuristic approach for filtering sequences. The results show that RAFTS3 is much faster than BLASTp with negligible loss of sensitivity when applied against large

databases in all tests performed and can be successfully used in several biological data-mining tasks.

IMPLEMENTATION

Since RAFTS3 deals with protein sequence comparison against protein databases, the first step to be considered is to set up the protein database into a specific RAFTS3 format. The formatting consists of two steps to be applied to each protein sequence within a FASTA file: a) the sequences must be indexed by a hash function and b) a binary amino acid co-occurrence matrix (BCOM) has to be assigned to each sequence to represent its contents.

When a formatted database is available, query searches can be performed. This process is also divided in two distinct steps: (1) the filtering of candidates, that selects sequences whose indexed k-mers are shared with the query sequence, and (2) the comparison of these candidates, that is done by means of the BCOM.

Database formatting process

The formatting process takes a FASTA database as input and creates a file comprising a hash table and the BCOM matrices for all sequences in the database. Aiming to improve access to the sequences, RAFTS3 also creates an index to allow direct access to each sequence in the FASTA file (Figure 1.A).

For each sequence in the database a set of k-mers is randomly selected and submitted to a hash function. The indexes are then stored into a hash table for fast selection of candidate for comparison. These indexes will permit further retrieval of any sequence in the database sharing a given k-mer. As default, 10 k-mers with lengths of 6 amino acid residues are selected per sequence.

The formatting process also involves a BCOM assignment to each sequence. The BCOM was designed to represent the sequences using few bytes of memory. Both the hash table and the BCOM matrices are stored in a common structure that is loaded

in RAM with the application aiming to minimize disk access when comparing sequences. The hash function and the BCOM structure will be detailed further.

Query sequence search

Searching is the goal step in RAFTS3. Its purpose is to retrieve similar sequences to a sequence of interest from a database. Also, it is desirable that the recovered sequences are ranked by their similarity with the query sequence. Searching involves two main steps: filtering and comparison.

In the filtering process, the search scope is reduced by selecting, through a hash table, only sequences containing common k-mers related to the query sequence. To perform a search based on a sequence of a given length n, hash indexes for all possible k-mers with length k are calculated by taking a sliding window that runs through the sequence from position 1 to $n-k+1$. The indexes generated for each k-mer are used to select the candidate sequences by consulting the hash table (Figure 1.B).

The comparison is performed with the candidate sequences based on their BCOM. The details of the comparison method will be discussed later (see Binary co-occurrence matrix (BCOM)). Alignments of the best results can also be done to confirm the results or to assign them to a well-established metric. The number of alignments can be customized by parameters; by default, a Smith-Waterman alignment [19] is performed only with the best stated result. As a measure of alignment quality, besides the alignment score, we calculate a relative score E (1) [20]:

$$E = \frac{alignment\ score\ of\ S_1\ with\ S_2}{alignment\ score\ of\ S_1\ with\ S_1} \qquad (1)$$

where S_1 and S_2 are protein sequences. E-values are also computed using Karlin Altschul statistics [21].

Hash function for candidate sequence selection

The hash function of RAFTS3 is an essential step in the filtering process and it is applied to both database and query. The recursive indexing technique (INREC) [22]

was used to assign a real number to a protein k-mer. INREC is a technique of dimensionality reduction and pattern recognition that uses a recursive process of a mathematical function to encapsulate, in a single number, the information that describes a pattern. Thereby, the indexes generated by similar sequences are equal or close to each other. The numbers generated by the INREC function are transformed in hash indexes H through the expression (2).

$$H = \mathrm{mod}(\mathrm{INREC}(k\text{-}mer) \times largenumber, dbsize) \qquad (2)$$

Where $largenumber$ is a value to express the decimal fraction of the INREC index as an integer number, and $dbsize$ defines size and spreading of the hash table. By using the hash table, sequences sharing the same INREC indexes are rapidly selected as candidate for comparison.

To apply the INREC algorithm, amino acid residues need to be converted to a quaternary numeral system triplet by a two-way conversion table (Table 1). The numbers are arbitrary, but the codes are assigned in correspondence to possible codons. The numerals 1, 2, 3 and 4 represent the nucleotide residues A, C, G and T/U, respectively.

Thus, given a sequence of integers $D = \{d_1, d_2, ..., d_m\}$ representing a sequence of length m, where $d_1 \in \{1, 2, 3, 4\}$. The INREC index I is generated from the recursion of the function f:

$$I = f(d_1 f(d_2 ... f(d_m))) \qquad (3)$$

where,

$$f(d_i) = tanh\left(\sqrt{\left(\frac{d_i}{4}\right)^{-1}}\right) \qquad (4)$$

The amino acid sequence MAF can be used to illustrate how the indexing works. By using the conversion table (Table 1) the amino acids are represented as $M = \{1, 4, 3\}$, $A = \{3, 2, 4\}$ and $F = \{4, 4, 2\}$. Thus, the sequence MAF can be

represented as $D = \{1,4,3,3,2,4,4,4,2\}$. Applying the f function recursively (4 and 3), from the last element to the first:

for $i = 9$, $d_i = 2$, $f(2) = 0.88$

for $i = 8$, $d_i = 4$, $f\left(4 \times f\left(2\right)\right) = 0.78$

for $i = 7$, $d_i = 4$, $f\left(4 \times f\left(4 \times f\left(2\right)\right)\right) = 0.81$

for $i = 6$, $d_i = 1$, $f\left(1 \times f\left(4 \times f\left(4 \times f\left(2\right)\right)\right)\right) = 0.80$

for $i = 5$, $d_i = 2$, $f\left(2 \times f\left(1 \times f\left(4 \times f\left(4 \times f\left(2\right)\right)\right)\right)\right) = 0.91$

for $i = 4$, $d_i = 3$, $f\left(3 \times f\left(2 \times f\left(1 \times f\left(4 \times f\left(4 \times f\left(2\right)\right)\right)\right)\right)\right) = 0.83$

for $i = 3$, $d_i = 3$, $f\left(3 \times f\left(3 \times f\left(2 \times f\left(1 \times f\left(4 \times f\left(4 \times f\left(2\right)\right)\right)\right)\right)\right)\right) = 0.85$

for $i = 2$, $d_i = 4$, $f\left(4 \times f\left(3 \times f\left(3 \times f\left(2 \times f\left(1 \times f\left(4 \times f\left(4 \times f\left(2\right)\right)\right)\right)\right)\right)\right)\right) = 0.79$

for $i = 1$, $d_i = 1$,

$$f\left(1 \times f\left(4 \times f\left(3 \times f\left(3 \times f\left(2 \times f\left(1 \times f\left(4 \times f\left(4 \times f\left(2\right)\right)\right)\right)\right)\right)\right)\right)\right) = 0.97$$

Therefore, for the sequence MAF, the INREC index I is 0.97.

Binary co-occurrence matrix (BCOM)

The binary co-occurrence matrix BCOM is a bi-dimensional fingerprint of an amino acid sequence. It not only represents an amino acid sequence but is a pattern for comparison with other sequences.

A BCOM is a binary matrix where each cell position (x, y) represents the occurrence of an amino acid pair XY in a sequence S. If the value within the cell is set to null, the pair does not occur in S (Figure 2). Thus for each sequence a 20x20 binary matrix is generated representing the occurrence of all possible amino acid pairs within it. Thereby, any sequence can be represented by a matrix with 400 bits or 50 bytes. The small data volume and the uniform structure of the BCOM allows databases

with millions of sequences to be represented and stored in RAM. The entire NR database can be handled in a common laptop.

To compare two matrices, let A and B be BCOMs corresponding to sequences S_1 and S_2 respectively. The binary sum between the matrices A and B represents the occurrence of common amino acid residue pairs and reflects the sequences similarity. Similarly, the binary operation xor is performed to calculate the degree of dissimilarity as a support for the comparison. Thus, the measure of difference e between A and B is given by the equation (5).

$$e = \frac{sum(xor(A,B))}{sum(and(A,B))} \qquad (5)$$

Each candidate sequence selected in the filter step is related to a dissimilarity measure given by e. Finally, correlation coefficients r (6) between the matrices are also calculated for BCOMs of sequences with highest similarity based on e and are used for reordering the results. Correlation coefficients are usually used to compare image differences; here the same was done with the BCOMs as an estimate of sequence identity. For instance, the sequence of the major facilitator superfamily protein of *Serratia* sp. *AS12* (gi 333925879) shares about 80% identity with the arabinose efflux permease family protein of *Rahnella aquatilis* (gi 383191252) and the correlation coefficient is 73%; in contrast the amino acid transporter of *Aspergillus oryzae* shares about 20% of identity with the former while the correlation coefficient is 28%.

$$r = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \bar{B})^2\right)}} \qquad (6)$$

Where, $\bar{A} = Mean(A)$ and $\bar{B} = Mean(B)$.

Due the computational cost, the number of sequences compared with the correlation equation (6) was limited to 50.

Implementation and datasets

RAFTS3 was written in MATLAB using its built-in functions, the Bioinformatics Toolbox [23] and an in-house library. Three protein databases were used, the NCBI NR with 19,689,576 sequences, PFAM [24] with 15,929,002 sequences and the UniRef50 [25] with 6,784,251 sequences. The performance and sensitivity of RAFTS3 was compared with that of BLASTp version 2.2.26+, USEARCH and PAUDA. Tests were performed using Linux CentOS 6.5 on a Desktop AMD Six-Core 3.5Ghz processor with 8Gb of RAM, configuration details for each test are explained on each results section.

RESULTS AND DISCUSSION

Parameters selection analysis

To determine the default parameters to be used by RAFTS3 for the candidate selection step, sets of 1 to 20 k-mers with 4, 5, 6 or 7 amino acid residues were evaluated using the NR database. A subset of 1000 protein sequences randomly selected from NR was used as query. Two criteria were considered to define the RAFTS3 configuration settings: the running time to search 1000 queries (Figure 3.A); and the number of queries with second best hit with relative score higher than 0.3 (Figure 3.B). The best hit was disregarded since that always corresponds to the query sequence.

The purpose of this procedure was to find the number and size of k-mers to be adopted as default parameters to carry out searches with RAFTS3. This analysis showed that the running times were lower using k-mer sizes of 6 and 7 residues and the number of hits with relative score higher than 0.3 reached a plateau with sets of 10 k-mers per sequence. Thereby, the following parameters were chosen as default: 10 k-mers of 6 amino acid residues per sequence.

Comparison of RAFTS3 with BLASTp

The sensitivity and running time of the RAFTS3 was compared with BLASTp using 1000 sequences randomly selected from a newer version of NR. These sequences were absent in the database used for search tests and represent sequences from more than 650 different organisms. This comparison simulates an automated annotation task, thus BLASTp and RAFTS3 were configured only to report the best hit for comparison. The sensitivity was evaluated as the number of similar sequences retrieved and the processing time spent in the search by both tools. The number of sequences retrieved by BLASTp was considered as the gold standard, representing 100% of the results.

RAFTS3 showed results from 77% to 95% of sensitivity compared with BLASTp when searching UniRef50 database, from 86% to 95% when searching the Pfam database and from 89% to 97% when searching the NR database, depending on the threshold of the score (Table 2). RAFTS3 showed to be more than 300 times faster than BLASTp when searching in the larger database.

To illustrate the differences between the RAFTS3 and BLASTp hits, three different proteins were searched against the NR database: the pyrR (UniProtAC P39765) of *Bacillus subtilis* that regulates the transcription of the pyrimidine nucleotide (pyr) operon; the PRNP (UniProtAC P04165) of *Homo sapiens* related with neuronal development and synaptic plasticity; and the PSG1 (UniProtAC P11464) of *Homo sapiens* related with female pregnancy. The top 10 hits found by each were selected and the E-value and the relative scores were calculated for comparison. The results showed that, despite some differences, RAFTS3 performed similarly to BLASTp (Table 3, 4 and 5).

To compare the ranking order of sequences given by BLASTp and RAFTS3, 1000 sequences were randomly selected from the dataset to be used as query against the NR database. The position of RAFTS3 best hits were scored among BLASTp top 50 hits and vice-versa. The results showed that 72% of the RAFTS3 best hits occurred within the first 10 BLASTp top results (Supplementary material Table S1), suggesting

that sequences retrieved by RAFTS3 are in most the same or very closely related to that retrieved by BLASTp. To better illustrate the ranking differences between BLASTp and RAFTS3 the top 50 hits identified by RAFTS3 and BLASTp using 5 different proteins randomly selected from the test set as query to search against the NR database are shown in supplementary material Table S2. In all cases, BLASTp best hit was among the 10 best hits of RAFTS3. Interestingly, for steroidogenic factor 1 isoform X2 RAFTS3 top hit had a higher relative score that of BLASTp (Table S2).

Comparison of RAFTS3 with USEARCH

USEARCH provides freely only a version with limited use of resources, the complete version of requires a paid license. Thereby, we chose to use the small COG [26] database to compare RAFTS3 and USEARCH performances. In this test, USEARCH was faster and more accurate than RAFTS3. However, due to the limitations of the free version, it was not possible to evaluate USEARCH performance searching large databases. It is possible to anticipate that memory consumption of USEARCH will be more than 40Gb for the NR database, while RAFTS3 uses 20 times less. Also RAFTS3 runtime is not much affected by the database size and the sensitivity tends to increase. These considerations indicate that the use of RAFTS3 may be advantageous over USEARCH when searching large databases.

Comparison of RAFTS3 with PAUDA

To compare RAFTS3 with PAUDA an executable was developed to translate DNA sequences in all 6 frames to search on a protein database. We called it RAFST3x (in analogy to BLASTx). The tests were performed comparing 1000 sequences randomly selected from the NT database (lengths from 50 to 3000 pb) against the UniRef50 database. RAFTS3x was 7% faster than PAUDA and more sensitive, yielding twice as many hits above the threshold relative score (Table 6).

Alignment information

The performance advantage of RAFTS3 relies on the comparison of sequences without the need of alignment. The measure used is based on the BCOM's comparison that have some relationship with an alignment score. It's possible to perform local alignments on the hits reported by RAFTS3 using the Smith-Waterman algorithm, however this adds an additional cost on time. The runtime for RAFTS3 configurations using from 0 to 100 alignments to search the 1000 sequences against the NR database varied from 40 seconds to 17 minutes. Thus this option must be used wisely.

CONCLUSIONS

RAFTS3 uses an aggressive filter approach with a fast comparison method based on BCOMs. Due to the limitation of the free version of USEARCH, comparisons for searches against large databases could not be performed. The comparison of RAFTS3 with BLASTp showed that RAFTS3 could be used to achieve fast protein similarity searches with a small loss of sensitivity. The sensitivity compared to BLASTp increases with the sequence similarity. RAFTS3 also shows a minimal loss on performance when challenged with larger databases in comparison with BLASTp, as judged by the increase in time to search on UniRef50 compared to NR (almost 3 times as large), the running time for RAFTS3 increased twice while BLASTp increased thrice. Thus RAFTS3 could be especially advantageous when using large databases with many sequences being queried. As the database increases, the filtering options can be made more stringent avoiding the increase of the number of candidate sequences selected and, consequently, of memory usage.

We have demonstrated that the RAFTS3 can perform high-speed protein search comparisons locally using a desktop computer or laptop. RAFTS3 is being used in tasks as genome annotation by our Bioinformatics group at the Federal University of Parana with success and presents a good solution for protein sequence data mining.

DECLARATIONS

Availability and requirements

Project name: RAFTS3

Project home page: https://sourceforge.net/projects/rafts3/

Operating system(s): Linux, Windows

Programming language: MATLAB (R2012a)

Other requirements: MCR MATLAB Compiler Runtime v7.17 (only for compiled version)

License: Source code and binaries freely available under the BSD License

Any restrictions to use by non-academics: None

Competing interests

The authors declare that they have no competing interests.

Funding

National Institute of Science and Technologies of Biological Nitrogen Fixation, Fundação Araucária, CAPES and CNPq.

Authors' contributions

RAV implemented the software, validated the results and wrote the manuscript. RTR designed the study and developed the prototype. FOP, VAW, JNM and EMS contributed to the concepts and revised the manuscript. DG contributed to the concepts. JHT contributed to the testing. All authors read and approved the final manuscript.

REFERENCES

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

2. Sayers EW, Barrett T, Benson D a, Bolton E, Bryant SH, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2011;39: D38–51. doi:10.1093/nar/gkq1172

3. Vinga S, Almeida J. Alignment-free sequence comparison--a review. Bioinformatics. 2003;19: 513–523. doi:10.1093/bioinformatics/btg005

4. Mahmood K, Webb GI, Song J, Whisstock JC, Konagurthu AS. Efficient large-scale protein sequence comparison and gene matching to identify orthologs and co-orthologs. Nucleic Acids Res. 2012;40: e44–e44. doi:10.1093/nar/gkr1261

5. Mantaci S, Restivo A, Sciortino M. Distance measures for biological sequences: Some recent approaches. Int J Approx Reason. 2008;47: 109–124. doi:10.1016/j.ijar.2007.03.011

6. Giancarlo R, Scaturro D, Utro F. Textual data compression in computational biology: a synopsis. Bioinformatics. 2009;25: 1575–1586. doi:10.1093/bioinformatics/btp117

7. Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci U S A. 1986;83: 5155–9. Available: http://www.ncbi.nlm.nih.gov/pubmed/3460087

8. Huang G, Zhou H, Li Y, Xu L. Alignment-free comparison of genome sequences by a new numerical characterization. J Theor Biol. 2011;281: 107–12. doi:10.1016/j.jtbi.2011.04.003

9. Yu C, Cheng S-Y, He RL, Yau SS-T. Protein map: an alignment-free sequence comparison method based on various properties of amino acids. Gene. 2011;486: 110–8. doi:10.1016/j.gene.2011.07.002

10. Comin M, Verzotto D. Alignment-free phylogeny of whole genomes using underlying subwords. Algorithms Mol Biol. 2012;7: 34. doi:10.1186/1748-7188-7-34

11. Soares I, Goios A, Amorim A. Sequence Comparison Alignment-Free Approach Based on Suffix Tree and L-Words Frequency. Sci World J. 2012;2012: 1–4. doi:10.1100/2012/450124

12. Kurtz S, Narechania A, Stein JC, Ware D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics. 2008;9: 517. doi:10.1186/1471-2164-9-517

13. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16: 236. doi:10.1186/s12864-015-1419-2

14. Kantorovitz MR, Robinson GE, Sinha S. A statistical method for alignment-free

comparison of regulatory sequences. Bioinformatics. 2007;23: i249–i255. doi:10.1093/bioinformatics/btm211

15. Pevzner PA. Statistical distance between texts and filtration methods in sequence comparison. Bioinformatics. 1992;8: 121–127. doi:10.1093/bioinformatics/8.2.121

16. Huson DH, Xie C. A poor man's BLASTX--high-throughput metagenomic protein database search using PAUDA. Bioinformatics. 2014;30: 38–39. doi:10.1093/bioinformatics/btt254

17. Kumar R, Mishra BK, Lahiri T, Kumar G, Kumar N, Gupta R, et al. PCV: An Alignment Free Method for Finding Homologous Nucleotide Sequences and its Application in Phylogenetic Study. Interdiscip Sci. 2016; doi:10.1007/s12539-015-0136-5

18. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26: 2460–2461. doi:10.1093/bioinformatics/btq461

19. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147: 195–197. doi:10.1016/0022-2836(81)90087-5

20. Barbosa-Silva A, Satagopam VP, Schneider R, Ortega JM. Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence. BMC Bioinformatics. 2008;9: 141. doi:10.1186/1471-2105-9-141

21. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A. 1990;87: 2264–8. Available: http://www.ncbi.nlm.nih.gov/pubmed/2315319

22. Souza JA. Reconhecimento de padrões usando indexação recursiva. Thesis. Universidade Federal de Santa Catarina. 1999. Available: https://repositorio.ufsc.br/handle/123456789/80484. Accessed 31 May 2016.

23. The MathWorks Inc. MATLAB and Bioinformatics Toolbox Release 2012b. Natick, Massachusetts, United States; Available: http://www.mathworks.com. Accessed 31 May 2016.

24. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H-R, et al. The Pfam protein families database. Nucleic Acids Res. 2007;36: D281–D288. doi:10.1093/nar/gkm960

25. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23: 1282–1288. doi:10.1093/bioinformatics/btm098

26. Tatusov RL, Koonin E V, Lipman DJ. A genomic perspective on protein families. Science. 1997;278: 631–7.

FIGURES

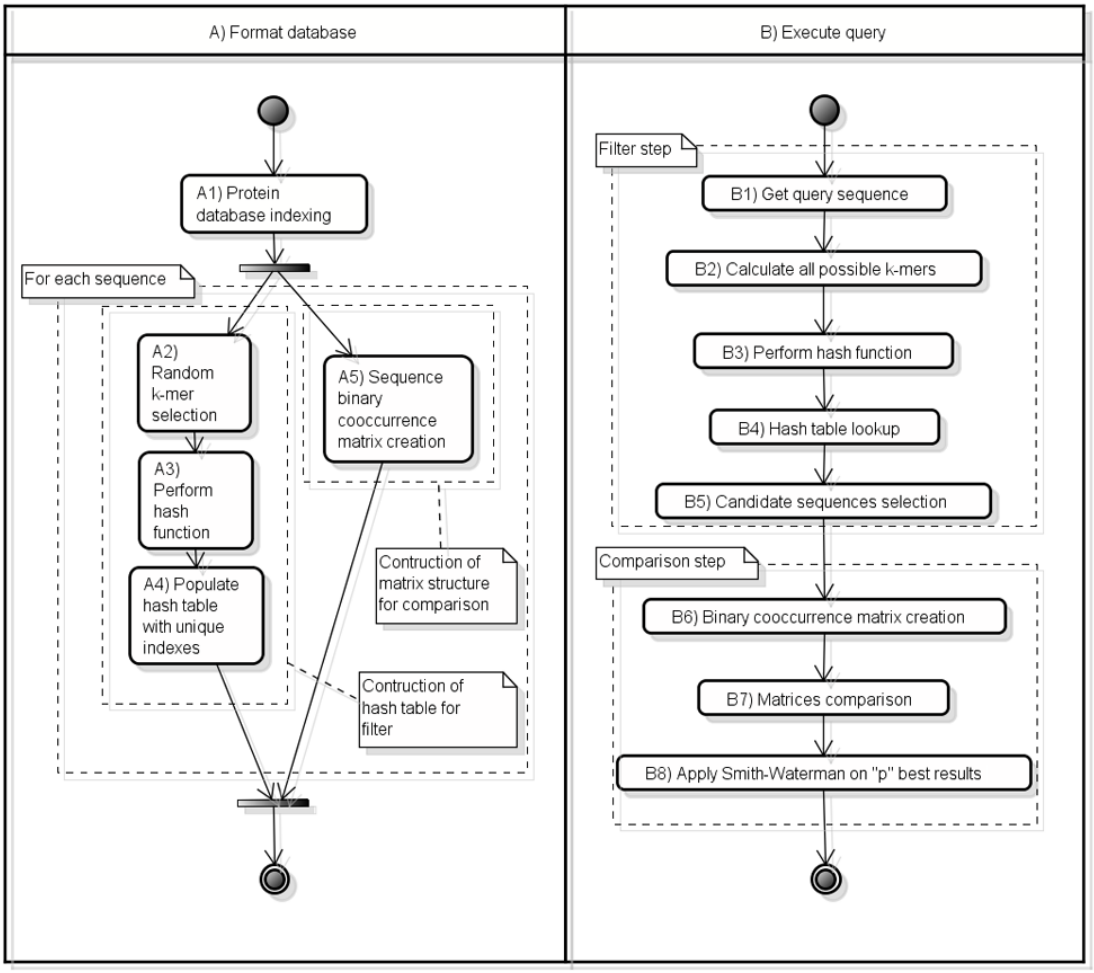

Figure 1. RAFTS3 activity diagram. RAFTS3 format database and query search overview. A) Shows the database formatting processes, which involve construction of two structures used in query sequence search, a hash table and a set of binary co-occurrence matrices. B) Shows the process for searching and comparison of a query sequence, with filtering and comparison steps separated.

Figure 2. Binary co-occurrence matrix (BCOM). Co-occurrence matrix of a protein sequence. White squares represent the occurrence of amino acid pairs; black squares represent non-occurrence.

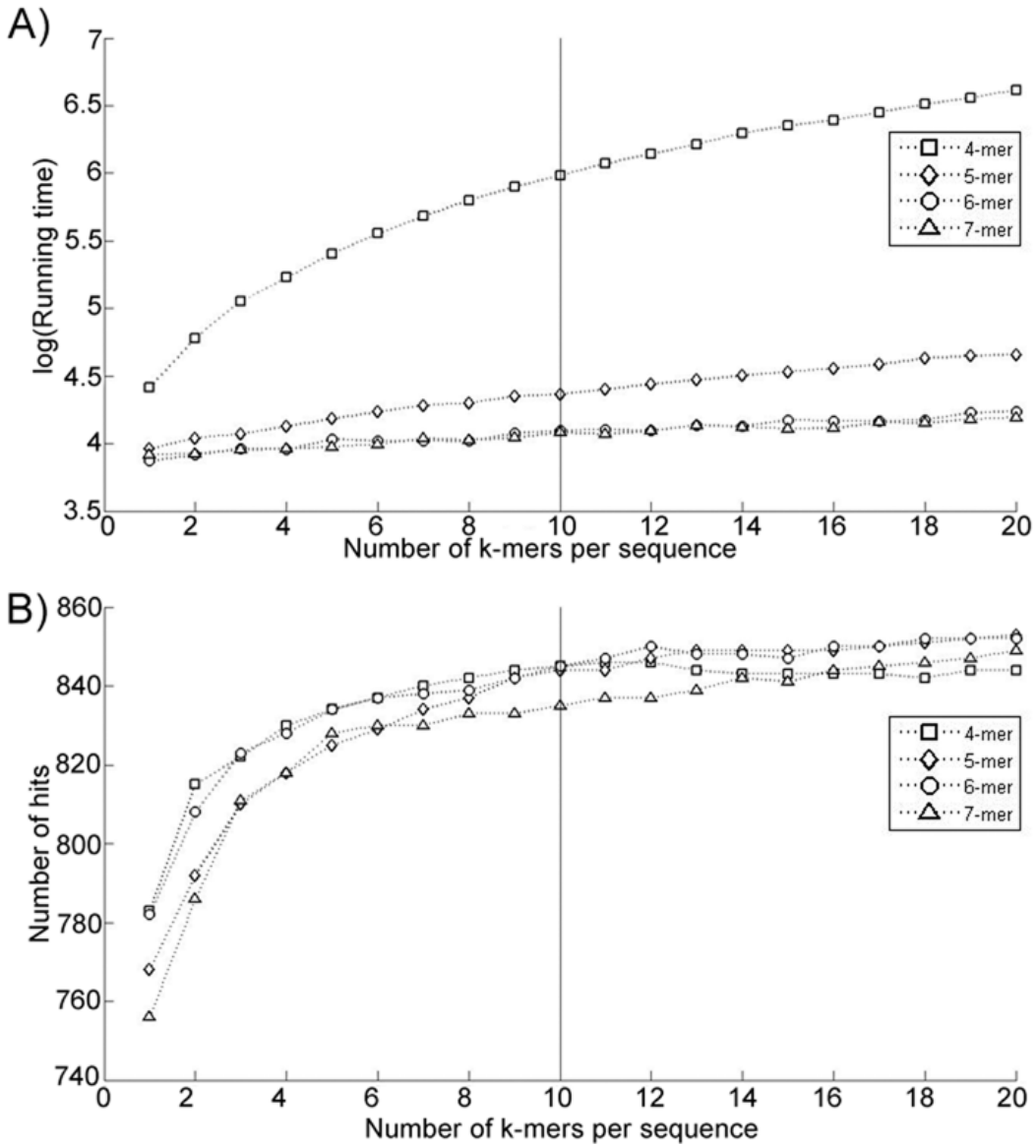

Figure 3. Parameters selection and configuration testing. Comparison of different k-mer sets. The comparison was made by analyzing the second best hit of the search results of 1000 sequences randomly selected. The number of k-mers ranged from 1 to 20 and their lengths were 4, 5, 6 and 7 amino acid residues. A) Shows the logarithm of the running time in seconds to search 1000 queries for each configuration. B) Shows the number of queries with second best hit with relative score over 0.3.

TABLES

Table 1. Amino acid numeric conversion

Amino acid	Code
A	3 2 4
R	1 3 1
N	1 1 2
D	3 1 2
C	4 3 2
Q	2 1 1
E	3 1 1
G	3 3 4
H	2 1 4
I	1 4 1
L	2 4 2
K	1 1 1
M	1 4 3
F	4 4 2
P	2 2 4
S	4 2 1
T	1 2 4
W	4 3 3
Y	4 1 2
V	3 4 2

Two-way conversion table of amino acid residues to quaternary numeral system triplets.

Table 2. Performance comparison of similarity search tools on the same query dataset (1000 sequences) against different protein databases

Database	Total sequences	Total aa	Running time		Percentage of sequences over relative score threshold found by RAFTS3 compared with BLASTp						
			BLASTp	RAFTS3	0.3	0.4	0.5	0.6	0.7	0.8	0.9
UniRef50	6,784,251	2,189,361,886	120m16.516s	0m47.209s	81%	80%	77%	80%	87%	92%	95%
Pfam	15,929,002	5,169,768,107	262m22.350s	0m52.832s	86%	91%	92%	93%	95%	95%	94%
NR	19,689,576	6,752,058,980	362m1.048s	1m7.514s	89%	92%	93%	95%	96%	97%	96%

The total number of sequences and amino acids included in each database are shown in the “Total sequences” and “Total aa” columns, respectively. The ratio RAFTS3/BLASTp gives the fraction of RAFTS3 hits over BLASTp hits for the indicated relative score thresholds.

Table 3. Comparison of top 10 of BLASTp and RAFTS3 hits of *Bacillus subtilis* PyrR protein

RAFTS3			**BLASTp**		
Query - UniProtAC: P39765 Bifunctional protein PyrR OS=*Bacillus subtilis*					
Relative Score	**E-Value**	**Subject Sequence**	**Relative Score**	**E-value**	**Subject Sequence**
1	1.6941e-50	gi|16078611|ref|NP_389430.1| bifunctional pyrimidine regulatory protein PyrR/uracil phosphoribosyltransferase [*Bacillus subtilis subsp. subtilis str. 168*]	1	1.6941e-50	gi|16078611|ref|NP_389430.1| bifunctional pyrimidine regulatory protein PyrR/uracil phosphoribosyltransferase [*Bacillus subtilis subsp. subtilis str. 168*]
0.991274	4.9554e-50	gi|296331123|ref|ZP_06873597.1| bifunctional pyrimidine regulatory protein PyrR uracil phosphoribosyltransferase [*Bacillus subtilis subsp. spizizenii ATCC 6633*]	0.991274	4.9554e-50	gi|296331123|ref|ZP_06873597.1| bifunctional pyrimidine regulatory protein PyrR uracil phosphoribosyltransferase [*Bacillus subtilis subsp. spizizenii ATCC 6633*]
0.958988	2.5129e-48	gi|1373160|gb|AAB57770.1| PyrR, partial [*Bacillus subtilis subsp. subtilis str. 168*]	0.973822	4.24e-49	gi|398304125|ref|ZP_10507711.1| bifunctional pyrimidine regulatory protein PyrR uracil phosphoribosyltransferase [*Bacillus vallismortis DV1-F-3*]
0.973822	4.24e-49	gi|398304125|ref|ZP_10507711.1| bifunctional pyrimidine regulatory protein PyrR uracil phosphoribosyltransferase [*Bacillus vallismortis DV1-F-3*]	0.954625	4.4966e-48	gi|154685963|ref|YP_001421124.1| bifunctional pyrimidine regulatory protein PyrR uracil phosphoribosyltransferase [*Bacillus amyloliquefaciens FZB42*]
0.954625	4.4966e-48	gi|154685963|ref|YP_001421124.1| bifunctional pyrimidine regulatory protein PyrR uracil phosphoribosyltransferase [*Bacillus amyloliquefaciens FZB42*]	0.951134	6.9078e-48	gi|311068068|ref|YP_003972991.1| bifunctional pyrimidine regulatory protein PyrR/uracil phosphoribosyltransferase [*Bacillus atrophaeus 1942*]
0.953752	5.006e-48	gi|375362191|ref|YP_005130230.1| bifunctional pyrimidine regulatory protein PyrR/uracil phosphoribosyltransferase *[Bacillus amyloliquefaciens subsp. plantarum CAU B946*]	0.958988	2.5129e-48	gi|1373160|gb|AAB57770.1| PyrR, partial [*Bacillus subtilis subsp. subtilis str. 168*]
0.951134	6.9078e-48	gi|311068068|ref|YP_003972991.1| bifunctional pyrimidine regulatory protein PyrR/uracil phosphoribosyltransferase [*Bacillus atrophaeus 1942*]	0.953752	5.006e-48	gi|375362191|ref|YP_005130230.1| bifunctional pyrimidine regulatory protein PyrR/uracil phosphoribosyltransferase [*Bacillus amyloliquefaciens subsp. plantarum CAU B946*]
0.904887	2.0525e-45	gi|52080149|ref|YP_078940.1| bifunctional pyrimidine regulatory protein PyrR uracil phosphoribosyltransferase [*Bacillus licheniformis DSM 13 = ATCC 14580*]	0.904887	2.0525e-45	gi|52080149|ref|YP_078940.1| bifunctional pyrimidine regulatory protein PyrR uracil phosphoribosyltransferase [*Bacillus licheniformis DSM 13 = ATCC 14580*]
0.877836	5.6563e-44	gi|389573331|ref|ZP_10163406.1| bifunctional protein PyrR [*Bacillus aerophilus KACC 16563*]	0.877836	5.6563e-44	gi|389573331|ref|ZP_10163406.1| bifunctional protein PyrR [*Bacillus aerophilus KACC 16563*]
0.873473	9.6739e-44	gi|157692227|ref|YP_001486689.1| pyrR gene product [*Bacillus pumilus SAFR-032*]	0.872600	1.077e-43	gi|194014677|ref|ZP_03053294.1| bifunctional protein PyrR [*Bacillus pumilus ATCC 7061*]

Comparison of the ten first results of RAFTS3 and BLASTp searching the *Bacillus subtilis* PyrR protein against the NR database. The subject sequences are ordered by each software default criteria.

Table 4. Comparison of top 10 of BLASTp and RAFTS3 hits of *Homo sapiens* Prion protein

RAFTS3			**BLASTp**		
Query - UniProtAC: P04156 Major prion protein OS=*Homo sapiens*					
Relative Score	**E-Value**	**Subject Sequence**	**Relative Score**	**E-value**	**Subject Sequence**
1	3.2879e-83	gi|4506113|ref|NP_000302.1| major prion protein preproprotein [*Homo sapiens*]	1	3.2879e-83	gi|4506113|ref|NP_000302.1| major prion protein preproprotein [*Homo sapiens*]
0.925566	8.6244e-77	gi|747847|emb|CAA58442.1| prion protein [*Homo sapiens*]	0.997843	5.0709e-83	gi|60834334|gb|AAX37089.1| prion protein [synthetic construct]
0.861920	3.1761e-71	gi|11128458|gb|AAC62750.2| prion protein precursor [*Homo sapiens*]	0.997843	5.0709e-83	gi|54695820|gb|AAV38282.1| prion protein (p27-30) (Creutzfeld-Jakob disease, Gerstmann-Strausler-Scheinker syndrome, fatal familial insomnia) [synthetic construct]
0.925566	8.6596e-77	gi|54695822|gb|AAV38283.1| prion protein (p27-30) (Creutzfeld-Jakob disease, Gerstmann-Strausler-Scheinker syndrome, fatal familial insomnia) [synthetic construct]	0.994606	1.091e-82	gi|397501420|ref|XP_003821383.1| PREDICTED: major prion protein isoform 6 [*Pan paniscus*]
0.896440	3.2079e-74	gi|38490002|gb|AAR21603.1| prion protein [*Homo sapiens*]	0.997843	5.0509e-83	gi|474359|gb|AAC50089.1| prion protein [*Gorilla gorilla*]
0.995146	8.6386e-83	gi|189053893|dbj|BAG35206.1| unnamed protein product [*Homo sapiens*]	0.996764	6.2604e-83	gi|15277486|gb|AAH12844.1| Prion protein [*Homo sapiens*]
0.987594	3.8205e-82	gi|123237246|emb|CAM27320.1| prion protein (p27-30) (Creutzfeldt-Jakob disease, Gerstmann-Strausler-Scheinker syndrome, fatal familial insomnia) [*Homo sapiens*]	0.996764	6.2604e-83	gi|89160954|gb|ABD63004.1| prion protein PrP [*Homo sapiens*]
0.997843	5.0709e-83	gi|54695820|gb|AAV38282.1| prion protein (p27-30) (Creutzfeld-Jakob disease, Gerstmann-Strausler-Scheinker syndrome, fatal familial insomnia) [synthetic construct]	0.994606	9.6174e-83	gi|57114055|ref|NP_001009093.1| major prion protein preproprotein [*Pan troglodytes*]
0.996764	6.2604e-83	gi|15277486|gb|AAH12844.1| Prion protein [*Homo sapiens*]	0.994067	1.0707e-82	gi|18490397|gb|AAH22532.1| Prion protein [*Homo sapiens*]
0.773463	1.1262e-63	gi|194381546|dbj|BAG58727.1| unnamed protein product [*Homo sapiens*]	0.992449	1.4775e-82	gi|54695862|gb|AAV38303.1| prion protein (p27-30) (Creutzfeld-Jakob disease, Gerstmann-Strausler-Scheinker syndrome, fatal familial insomnia) [*Homo sapiens*]

Comparison of the ten first results of RAFTS3 and BLASTp searching the *Homo sapiens* Prion protein against the NR database. The subject sequences are ordered by each software default criteria.

Table 5. Comparison of top 10 of BLASTp and RAFTS3 hits of *Homo sapiens* PSG1 protein

RAFTS3			**BLASTp**		
Query - UniProtAC: P11464 Pregnancy-specific beta-1-glycoprotein 1 OS=*Homo sapiens*					
Relative Score	**E-Value**	**Subject Sequence**	**Relative Score**	**E-value**	**Subject Sequence**
1	3.9757e-128	gi|296317345|ref|NP_001171754.1| pregnancy-specific beta-1-glycoprotein 1 isoform 2 precursor [*Homo sapiens*]	1	3.9757e-128	gi|296317345|ref|NP_001171754.1| pregnancy-specific beta-1-glycoprotein 1 isoform 2 precursor [*Homo sapiens*]
0.996463	1.1629e-127	gi|190645|gb|AAA36515.1| pregnancy-specific glycoprotein-1a [*Homo sapiens*]	0.996463	1.1629e-127	gi|190645|gb|AAA36515.1| pregnancy-specific glycoprotein-1a [*Homo sapiens*]
0.985497	3.31e-126	gi|306797|gb|AAA52602.1| pregnancy-specific beta-glycoprotein c [*Homo sapiens*]	0.985497	3.2249e-126	gi|296317348|ref|NP_001171755.1| pregnancy-specific beta-1-glycoprotein 1 isoform 3 precursor [*Homo sapiens*]
0.985497	3.2249e-126	gi|296317348|ref|NP_001171755.1| pregnancy-specific beta-1-glycoprotein 1 isoform 3 precursor [*Homo sapiens*]	0.985497	3.31e-126	gi|306797|gb|AAA52602.1| pregnancy-specific beta-glycoprotein c [*Homo sapiens*]
0.981606	1.0502e-125	gi|306791|gb|AAA52590.1| pregnancy-specific beta-1-glycoprotein [*Homo sapiens*]	0.989034	1.1263e-126	gi|21361392|ref|NP_008836.2| pregnancy-specific beta-1-glycoprotein 1 isoform 1 precursor [*Homo sapiens*]
0.989034	1.1263e-126	gi|21361392|ref|NP_008836.2| pregnancy-specific beta-1-glycoprotein 1 isoform 1 precursor [*Homo sapiens*]	0.981606	1.0502e-125	gi|306791|gb|AAA52590.1| pregnancy-specific beta-1-glycoprotein [*Homo sapiens*]
0.985497	3.2945e-126	gi|190653|gb|AAA36517.1| pregnancy-specific glycoprotein-1d [*Homo sapiens*]	0.985497	3.2945e-126	gi|190653|gb|AAA36517.1| pregnancy-specific glycoprotein-1d [*Homo sapiens*]
0.985143	3.6678e-126	gi|190591|gb|AAA36511.1| pregnancy-specific beta-1-glycoprotein [*Homo sapiens*]	0.985143	3.6678e-126	gi|190591|gb|AAA36511.1| pregnancy-specific beta-1-glycoprotein [*Homo sapiens*]
0.569155	1.8476e-71	gi|3287447|gb|AAC25485.1| PSGIIA-a [*Homo sapiens*]	0.948709	2.2829e-121	gi|14250018|gb|AAH08405.1| PSG4 protein [*Homo sapiens*]
0.948709	2.2829e-121	gi|14250018|gb|AAH08405.1| PSG4 protein [*Homo sapiens*]	0.949416	1.8419e-121	gi|332855947|ref|XP_512709.3| PREDICTED: pregnancy-specific beta-1-glycoprotein 1 isoform 3 [*Pan troglodytes*]

Comparison of the ten first results of RAFTS3 and BLASTp searching the *Homo sapiens* PSG1 protein against the NR database. The subject sequences are ordered by each software default criteria.

Table 6. Performance comparison of RAFTS3 and PAUDA

Software	Runtime	Number of hits per relative score threshold						
		0.3	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
RAFTS3	428s	665	512	435	358	282	211	125
PAUDA	458s	313	201	131	89	60	46	36

The number of hits represent the number of sequences retrieved with relative score higher than the relative score threshold.

SUPPORTING INFORMATION

Table S1. Comparison of BLASTp and RAFTS3 hits ranking.

Rank position	BLASTp best hits		RAFTS3 best hits	
	Number of best hits[A]	Mean of scores[B]	Number of best hits[C]	Mean of scores[B]
1	429	0.89	429	0.89
2	105	0.91	118	0.86
3	44	0.83	65	0.84
4	34	0.90	35	0.83
5	25	0.86	25	0.81
6	14	0.83	16	0.74
7	10	0.89	9	0.84
8	9	0.88	13	0.78
9	8	0.87	10	0.81
10	8	0.86	7	0.66
11	3	0.77	2	0.72
12	2	0.95	5	0.57
13	3	0.74	6	0.71
14	8	0.75	8	0.87
15	1	0.77	6	0.70
16	5	0.89	3	0.47
17	2	0.91	2	0.63
18	1	0.94	2	0.63
19	3	0.60	3	0.69
20	1	1.00	1	0.59
21	1	0.81	2	0.42
22	1	0.99	2	0.68
23	2	0.79	0	0.00
24	0	0.00	2	0.87
25	1	1.00	3	0.91
26	0	0.00	1	0.97
27	1	0.98	2	0.36
28	2	0.73	1	0.99
29	0	0.00	1	0.99
30	0	0.00	2	0.74
31	0	0.00	3	0.46
32	1	0.80	1	0.96
33	0	0.00	2	0.75
34	0	0.00	1	0.57
35	0	0.00	1	0.96
36	3	0.97	0	0.00
37	0	0.00	2	0.96
38	0	0.00	0	0.00
39	1	0.95	0	0.00
40	1	0.98	0	0.00
41	0	0.00	0	0.00
42	0	0.00	2	0.78
43	0	0.00	0	0.00
44	0	0.00	4	0.68
45	0	0.00	0	0.00
46	0	0.00	1	0.53
47	0	0.00	1	0.92
48	0	0.00	0	0.00
49	0	0.00	1	0.53
50	0	0.00	0	0.00

The number of best hits of each tool is located in the top 50 hits of the other. 1000 sequences of the data set were used as queries for searching NR database with BLASTp or RAFTS3, and the number of best hit of each tool was scored for each rank position of the other tool. For example, 429 best hits of BLASTp are also best hits of RAFTS3, while 105 best hits of RAFTS3 are second best hits of BLASTp.

[A] Number of RAFTS3 best hits occurring in the indicated BLASTp rank position.

[B] Average relative score of the best hits occurring in the indicated rank position.

[C] Number of BLASTp best hits occurring in the indicated RAFTS3 rank position.

Table S2. Comparison of top 50 results of BLASTp and RAFTS3 for 5 random proteins.

Query	Rank	BLASTp			RAFTS3		
		Score	E-value	GI	Score	E-value	GI
hypothetical protein [Nocardiopsis prasina]	1	0.72	4.63E-36	54026316	0.68	1.41E-33	108799494
hypothetical protein [Nocardiopsis prasina]	2	0.71	1.49E-35	300786153	0.68	7.40E-34	126435148
hypothetical protein [Nocardiopsis prasina]	3	0.68	7.40E-34	126435148	0.72	4.63E-36	54026316
hypothetical protein [Nocardiopsis prasina]	4	0.68	8.82E-34	379709308	0.71	4.91E-35	386772711
hypothetical protein [Nocardiopsis prasina]	5	0.68	1.41E-33	108799494	0.70	5.52E-35	379736307
hypothetical protein [Nocardiopsis prasina]	6	0.71	4.91E-35	386772711	0.70	8.44E-35	319949673
hypothetical protein [Nocardiopsis prasina]	7	0.68	1.90E-33	363419515	0.71	1.49E-35	300786153
hypothetical protein [Nocardiopsis prasina]	8	0.67	5.49E-33	312139946	0.67	7.01E-33	379748180
hypothetical protein [Nocardiopsis prasina]	9	0.70	9.30E-35	325964444	0.66	1.22E-32	379755468
hypothetical protein [Nocardiopsis prasina]	10	0.70	8.44E-35	319949673	0.70	9.30E-35	325964444
hypothetical protein [Nocardiopsis prasina]	11	0.66	1.30E-32	148271386	0.67	7.15E-33	254821758
hypothetical protein [Nocardiopsis prasina]	12	0.66	2.05E-32	342858396	0.66	1.88E-32	379763014
hypothetical protein [Nocardiopsis prasina]	13	0.70	5.52E-35	379736307	0.68	1.90E-33	363419515
hypothetical protein [Nocardiopsis prasina]	14	0.69	1.74E-34	302526736	0.68	8.82E-34	379709308
hypothetical protein [Nocardiopsis prasina]	15	0.65	5.16E-32	239985975	0.69	1.74E-34	302526736
hypothetical protein [Nocardiopsis prasina]	16	0.67	4.19E-33	378816503	0.67	4.19E-33	378816503
hypothetical protein [Nocardiopsis prasina]	17	0.66	1.22E-32	379755468	0.67	3.58E-33	397679577
hypothetical protein [Nocardiopsis prasina]	18	0.64	1.93E-31	120405695	0.67	3.54E-33	392136326
hypothetical protein [Nocardiopsis prasina]	19	0.66	1.88E-32	379763014	0.66	1.30E-32	148271386
hypothetical protein [Nocardiopsis prasina]	20	0.68	6.58E-34	382944866	0.61	9.04E-30	359774163
hypothetical protein [Nocardiopsis prasina]	21	0.68	7.32E-34	382944705	0.60	1.33E-29	385651505
hypothetical protein [Nocardiopsis prasina]	22	0.68	8.77E-34	392068192	0.68	6.58E-34	382944866
hypothetical protein [Nocardiopsis prasina]	23	0.68	8.64E-34	392185753	0.68	7.32E-34	382944705
hypothetical protein [Nocardiopsis prasina]	24	0.63	8.40E-31	182440470	0.67	5.49E-33	312139946
hypothetical protein [Nocardiopsis prasina]	25	0.67	2.96E-33	363999376	0.69	4.24E-34	374610654
hypothetical protein [Nocardiopsis prasina]	26	0.67	3.54E-33	392136326	0.64	1.60E-31	333989082
hypothetical protein [Nocardiopsis prasina]	27	0.67	3.58E-33	397679577	0.43	1.22E-20	377569218
hypothetical protein [Nocardiopsis prasina]	28	0.63	6.21E-31	311741875	0.63	6.21E-31	311741875
hypothetical protein [Nocardiopsis prasina]	29	0.67	7.01E-33	379748180	0.68	8.77E-34	392068192
hypothetical protein [Nocardiopsis prasina]	30	0.67	7.15E-33	254821758	0.67	5.23E-33	325674171
hypothetical protein [Nocardiopsis prasina]	31	0.67	5.23E-33	325674171	0.54	2.58E-26	392847751
hypothetical protein [Nocardiopsis prasina]	32	0.65	4.77E-32	296164157	0.62	1.33E-30	118472651

hypothetical protein [Nocardiopsis prasina]	33	0.61	9.04E-30	359774163	0.55	1.78E-26	365870395
hypothetical protein [Nocardiopsis prasina]	34	0.59	4.46E-29	333022807	0.55	1.72E-26	392086623
hypothetical protein [Nocardiopsis prasina]	35	0.59	1.10E-28	328880653	0.55	7.93E-27	358003015
hypothetical protein [Nocardiopsis prasina]	36	0.53	7.08E-26	331696237	0.55	1.77E-26	386691466
hypothetical protein [Nocardiopsis prasina]	37	0.65	5.36E-32	145222559	0.53	1.41E-25	163857205
hypothetical protein [Nocardiopsis prasina]	38	0.68	9.12E-34	375137662	0.50	3.53E-24	344998040
hypothetical protein [Nocardiopsis prasina]	39	0.57	5.95E-28	326330248	0.65	5.36E-32	145222559
hypothetical protein [Nocardiopsis prasina]	40	0.58	4.87E-28	354570978	0.05	2.07E+01	395776055
hypothetical protein [Nocardiopsis prasina]	41	0.62	1.05E-30	302523345	0.04	4.45E+01	386354368
hypothetical protein [Nocardiopsis prasina]	42	0.58	3.48E-28	392383971	0.05	1.56E+01	375143224
hypothetical protein [Nocardiopsis prasina]	43	0.63	2.90E-31	262204158	0.05	1.64E+01	83716891
hypothetical protein [Nocardiopsis prasina]	44	0.57	6.72E-28	337267501	0.05	2.00E+01	257068192
hypothetical protein [Nocardiopsis prasina]	45	0.57	1.28E-27	392849102	0.05	1.66E+01	167577907
hypothetical protein [Nocardiopsis prasina]	46	0.56	2.16E-27	387970235	0.05	2.42E+01	241206048
hypothetical protein [Nocardiopsis prasina]	47	0.56	2.19E-27	392520598	0.06	1.34E+01	218893839
hypothetical protein [Nocardiopsis prasina]	48	0.56	2.71E-27	393170990	0.06	6.61E+00	373478940
hypothetical protein [Nocardiopsis prasina]	49	0.56	2.40E-27	397688108	0.05	2.89E+01	359149253
hypothetical protein [Nocardiopsis prasina]	50	0.55	7.93E-27	358003015	0.05	4.14E+01	168009884
transposase [Bacillus cereus]	1	0.93	1.25E-79	206973981	0.91	1.97E-78	218896864
transposase [Bacillus cereus]	2	0.92	4.07E-79	75763559	0.86	8.35E-74	229085868
transposase [Bacillus cereus]	3	0.91	1.97E-78	218896864	0.92	1.48E-78	229090883
transposase [Bacillus cereus]	4	0.92	9.60E-79	391290908	0.93	1.25E-79	206973981
transposase [Bacillus cereus]	5	0.92	1.07E-78	206973765	0.91	3.88E-78	196038692
transposase [Bacillus cereus]	6	0.92	1.48E-78	229090883	0.91	4.81E-78	196038713
transposase [Bacillus cereus]	7	0.91	3.88E-78	196038692	0.92	4.07E-79	75763559
transposase [Bacillus cereus]	8	0.91	4.81E-78	196038713	0.92	1.07E-78	206973765
transposase [Bacillus cereus]	9	0.90	4.52E-77	196042445	0.90	4.52E-77	196042445
transposase [Bacillus cereus]	10	0.90	5.13E-77	75760431	0.83	2.22E-71	222094044
transposase [Bacillus cereus]	11	0.88	1.76E-75	206973407	0.88	1.76E-75	206973407
transposase [Bacillus cereus]	12	0.86	8.35E-74	229085868	0.92	9.60E-79	391290908
transposase [Bacillus cereus]	13	0.84	7.58E-72	228905309	0.84	7.58E-72	228905309
transposase [Bacillus cereus]	14	0.83	2.22E-71	222094044	0.82	4.48E-70	229073617
transposase [Bacillus cereus]	15	0.83	8.95E-71	229172505	0.82	2.91E-70	229095554
transposase [Bacillus cereus]	16	0.82	2.91E-70	229095554	0.83	8.95E-71	229172505
transposase [Bacillus cereus]	17	0.82	4.48E-70	229073617	0.90	5.13E-77	75760431
transposase [Bacillus cereus]	18	0.80	3.28E-68	229182210	0.80	3.28E-68	229182210
transposase [Bacillus cereus]	19	0.66	1.14E-55	225871669	0.55	2.44E-46	75762371
transposase [Bacillus cereus]	20	0.65	3.71E-55	227811612	0.52	8.47E-44	228911455
transposase [Bacillus cereus]	21	0.65	8.75E-55	254762474	0.65	3.71E-55	227811612
transposase [Bacillus cereus]	22	0.57	5.32E-48	301068226	0.66	1.14E-55	225871669
transposase [Bacillus cereus]	23	0.55	2.44E-46	75762371	0.65	8.75E-55	254762474
transposase [Bacillus cereus]	24	0.56	1.46E-46	228970158	0.47	2.87E-39	75764333
transposase [Bacillus cereus]	25	0.52	8.47E-44	228911455	0.57	5.32E-48	301068226

transposase [Bacillus cereus]	26	0.55	1.56E-45	229106961	0.46	2.25E-38	75759724
transposase [Bacillus cereus]	27	0.54	2.41E-45	229119272	0.44	1.68E-36	75763688
transposase [Bacillus cereus]	28	0.54	4.12E-45	206973911	0.56	1.46E-46	228970158
transposase [Bacillus cereus]	29	0.54	4.09E-45	221642251	0.47	2.33E-38	10956343
transposase [Bacillus cereus]	30	0.50	5.62E-41	228924890	0.47	5.40E-39	301068223
transposase [Bacillus cereus]	31	0.48	2.69E-40	228906894	0.47	9.24E-39	165873444
transposase [Bacillus cereus]	32	0.47	2.87E-39	75764333	0.47	1.27E-38	254739166
transposase [Bacillus cereus]	33	0.46	2.25E-38	75759724	0.04	1.10E+01	371777874
transposase [Bacillus cereus]	34	0.44	1.68E-36	75763688	0.04	1.42E+01	163786962
transposase [Bacillus cereus]	35	0.47	1.10E-38	47568876	0.03	3.00E+01	229580250
transposase [Bacillus cereus]	36	0.47	5.40E-39	301068223	0.03	2.98E+01	399003489
transposase [Bacillus cereus]	37	0.47	2.33E-38	10956343	0.03	7.00E+01	336315012
transposase [Bacillus cereus]	38	0.47	9.24E-39	165873444	0.03	6.54E+01	397602092
transposase [Bacillus cereus]	39	0.47	1.27E-38	254739166	0.03	1.15E+02	328859802
transposase [Bacillus cereus]	40	0.42	2.03E-34	10956376	0.04	1.13E+01	355670767
transposase [Bacillus cereus]	41	0.45	1.15E-36	23099091	0.03	4.22E+01	255954341
transposase [Bacillus cereus]	42	0.40	3.14E-33	254687682	0.04	1.66E+01	390944355
transposase [Bacillus cereus]	43	0.44	3.77E-36	383438775	0.03	3.86E+01	375163697
transposase [Bacillus cereus]	44	0.44	1.49E-35	386712685	0.03	4.04E+01	284035464
transposase [Bacillus cereus]	45	0.44	3.77E-36	52078969	0.03	7.45E+01	325912143
transposase [Bacillus cereus]	46	0.44	5.20E-36	383439073	0.03	4.78E+01	218192290
transposase [Bacillus cereus]	47	0.41	1.86E-33	261409030	0.03	1.32E+02	301776841
transposase [Bacillus cereus]	48	0.41	2.26E-33	315647012	0.04	1.79E+01	340380971
transposase [Bacillus cereus]	49	0.40	1.74E-32	261409860	0.03	1.06E+02	296278265
transposase [Bacillus cereus]	50	0.41	2.86E-33	372455285	0.03	5.29E+01	12697963
iroE [Klebsiella pneumoniae]	1	1.00	1.58E-82	262044290	1.00	1.58E-82	262044290
iroE [Klebsiella pneumoniae]	2	0.99	3.00E-82	397743876	0.99	1.35E-81	238894719
iroE [Klebsiella pneumoniae]	3	0.99	5.13E-82	386034811	0.99	3.00E-82	397743876
iroE [Klebsiella pneumoniae]	4	0.99	5.71E-82	397345874	0.99	5.13E-82	386034811
iroE [Klebsiella pneumoniae]	5	0.99	5.10E-82	152970230	0.99	5.10E-82	152970230
iroE [Klebsiella pneumoniae]	6	0.99	5.67E-82	397446209	0.99	8.72E-82	365141280
iroE [Klebsiella pneumoniae]	7	0.99	1.35E-81	238894719	0.99	5.71E-82	397345874
iroE [Klebsiella pneumoniae]	8	0.99	8.72E-82	365141280	0.99	5.67E-82	397446209
iroE [Klebsiella pneumoniae]	9	0.87	1.02E-71	288935507	0.87	1.13E-71	290509545
iroE [Klebsiella pneumoniae]	10	0.87	1.13E-71	290509545	0.86	4.11E-71	206579000
iroE [Klebsiella pneumoniae]	11	0.86	4.11E-71	206579000	0.87	1.02E-71	288935507
iroE [Klebsiella pneumoniae]	12	0.52	2.77E-41	378978774	0.49	1.37E-38	375002495
iroE [Klebsiella pneumoniae]	13	0.53	4.53E-42	376400045	0.48	2.19E-38	353606873
iroE [Klebsiella pneumoniae]	14	0.52	4.80E-41	375261636	0.49	5.45E-39	366059620
iroE [Klebsiella pneumoniae]	15	0.52	4.80E-41	397658746	0.48	2.97E-38	16761559
iroE [Klebsiella pneumoniae]	16	0.52	3.64E-41	376395752	0.51	1.79E-40	261341410
iroE [Klebsiella pneumoniae]	17	0.51	1.79E-40	261341410	0.49	1.28E-38	204929666
iroE [Klebsiella pneumoniae]	18	0.51	2.40E-40	376385392	0.49	9.29E-39	353661331
iroE [Klebsiella pneumoniae]	19	0.51	2.33E-40	157145930	0.48	3.75E-38	363551476
iroE [Klebsiella pneumoniae]	20	0.50	9.85E-40	295096492	0.49	1.28E-38	238909547
iroE [Klebsiella pneumoniae]	21	0.50	5.09E-40	376383330	0.50	9.85E-40	295096492
iroE [Klebsiella pneumoniae]	22	0.50	1.00E-39	376383956	0.49	1.59E-38	168238684
iroE [Klebsiella pneumoniae]	23	0.49	3.80E-39	397167683	0.47	2.88E-37	353596784
iroE [Klebsiella pneumoniae]	24	0.47	5.58E-37	317053692	0.49	6.05E-39	366083251
iroE [Klebsiella pneumoniae]	25	0.49	5.45E-39	366059620	0.47	1.69E-37	392819761
iroE [Klebsiella pneumoniae]	26	0.49	6.05E-39	322614400	0.48	2.19E-38	168823183

iroE [Klebsiella pneumoniae]	27	0.48	5.04E-38	365850221	0.48	5.73E-38	205353732
iroE [Klebsiella pneumoniae]	28	0.49	6.05E-39	366083251	0.03	1.83E+01	264676321
iroE [Klebsiella pneumoniae]	29	0.49	9.29E-39	353661331	0.49	1.28E-38	197249038
iroE [Klebsiella pneumoniae]	30	0.48	3.36E-38	336247161	0.03	6.87E+01	237799500
iroE [Klebsiella pneumoniae]	31	0.49	9.29E-39	168262124	0.50	1.00E-39	376383956
iroE [Klebsiella pneumoniae]	32	0.49	4.88E-39	168233729	0.04	6.28E+00	330817286
iroE [Klebsiella pneumoniae]	33	0.49	1.37E-38	375002495	0.04	3.39E+00	294010958
iroE [Klebsiella pneumoniae]	34	0.49	1.28E-38	204929666	0.50	2.35E-39	157418222
iroE [Klebsiella pneumoniae]	35	0.49	1.28E-38	238909547	0.43	1.99E-33	213424548
iroE [Klebsiella pneumoniae]	36	0.49	1.77E-38	353617579	0.04	8.44E+00	383777638
iroE [Klebsiella pneumoniae]	37	0.49	1.28E-38	197249038	0.49	5.56E-39	386598810
iroE [Klebsiella pneumoniae]	38	0.49	1.59E-38	168238684	0.03	4.36E+01	325271609
iroE [Klebsiella pneumoniae]	39	0.48	2.19E-38	168823183	0.03	2.79E+01	386818396
iroE [Klebsiella pneumoniae]	40	0.48	3.03E-38	56414734	0.03	1.85E+02	188534867
iroE [Klebsiella pneumoniae]	41	0.50	1.11E-39	338767125	0.03	6.84E+01	385653217
iroE [Klebsiella pneumoniae]	42	0.50	1.11E-39	222104800	0.03	2.31E+01	167829434
iroE [Klebsiella pneumoniae]	43	0.48	5.76E-38	379050444	0.03	5.08E+02	303327182
iroE [Klebsiella pneumoniae]	44	0.48	2.19E-38	353606873	0.03	4.99E+01	328770351
iroE [Klebsiella pneumoniae]	45	0.48	2.97E-38	16761559	0.04	1.21E+01	87309978
iroE [Klebsiella pneumoniae]	46	0.48	5.73E-38	205353732	0.05	4.74E+00	344172297
iroE [Klebsiella pneumoniae]	47	0.48	6.10E-38	375124587	0.05	4.12E+00	323358023
iroE [Klebsiella pneumoniae]	48	0.48	3.75E-38	363551476	0.04	3.02E+01	254413004
iroE [Klebsiella pneumoniae]	49	0.48	3.75E-38	353570067	0.04	1.90E+01	227875198
iroE [Klebsiella pneumoniae]	50	0.48	4.62E-38	198243714	0.04	1.14E+01	152968329
secernin-3 [Myotis lucifugus]	1	0.93	6.08E-121	301769725	0.91	1.71E-117	344268354
secernin-3 [Myotis lucifugus]	2	0.93	1.04E-120	281348304	0.92	8.90E-120	194222320
secernin-3 [Myotis lucifugus]	3	0.92	8.90E-120	194222320	0.91	2.23E-118	296490700
secernin-3 [Myotis lucifugus]	4	0.93	4.68E-120	57110813	0.91	4.24E-118	115495695
secernin-3 [Myotis lucifugus]	5	0.91	2.23E-118	296490700	0.93	1.04E-120	281348304
secernin-3 [Myotis lucifugus]	6	0.91	4.24E-118	115495695	0.93	6.08E-121	301769725
secernin-3 [Myotis lucifugus]	7	0.91	1.71E-117	344268354	0.93	4.68E-120	57110813
secernin-3 [Myotis lucifugus]	8	0.90	3.83E-116	109100111	0.90	3.83E-116	109100111
secernin-3 [Myotis lucifugus]	9	0.89	1.41E-115	296204486	0.89	1.53E-115	380791835
secernin-3 [Myotis lucifugus]	10	0.89	2.97E-115	38504671	0.89	2.97E-115	38504671
secernin-3 [Myotis lucifugus]	11	0.89	1.53E-115	380791835	0.89	9.73E-115	111601409
secernin-3 [Myotis lucifugus]	12	0.89	2.02E-115	332209358	0.87	3.13E-112	395837305
secernin-3 [Myotis lucifugus]	13	0.89	5.97E-115	397507616	0.87	2.05E-112	348585759
secernin-3 [Myotis lucifugus]	14	0.89	1.12E-115	291391765	0.89	1.08E-114	114581821
secernin-3 [Myotis lucifugus]	15	0.89	9.73E-115	111601409	0.89	2.02E-115	332209358
secernin-3 [Myotis lucifugus]	16	0.89	1.08E-114	114581821	0.87	1.83E-112	351715131
secernin-3 [Myotis lucifugus]	17	0.87	3.13E-112	395837305	0.89	5.97E-115	397507616
secernin-3 [Myotis lucifugus]	18	0.87	1.83E-112	351715131	0.89	1.41E-115	296204486
secernin-3 [Myotis lucifugus]	19	0.87	2.05E-112	348585759	0.89	1.12E-115	291391765
secernin-3 [Myotis lucifugus]	20	0.84	5.68E-109	297264350	0.84	2.58E-108	302058287
secernin-3 [Myotis lucifugus]	21	0.84	2.58E-108	302058287	0.84	3.17E-108	332209360
secernin-3 [Myotis lucifugus]	22	0.84	4.91E-108	397507618	0.84	5.68E-109	297264350
secernin-3 [Myotis lucifugus]	23	0.84	3.17E-108	332209360	0.83	9.35E-108	332814760
secernin-3 [Myotis lucifugus]	24	0.83	1.44E-107	354505419	0.84	4.91E-108	397507618
secernin-3 [Myotis lucifugus]	25	0.83	9.35E-108	332814760	0.83	1.44E-107	354505419
secernin-3 [Myotis lucifugus]	26	0.80	1.32E-103	61969660	0.80	1.32E-103	61969660
secernin-3 [Myotis lucifugus]	27	0.80	1.07E-103	149022245	0.80	2.79E-103	15929748

secernin-3 [Myotis lucifugus]	28	0.80	2.79E-103	15929748	0.72	9.78E-93	119631548
secernin-3 [Myotis lucifugus]	29	0.80	1.00E-102	74153182	0.80	1.00E-102	74153182
secernin-3 [Myotis lucifugus]	30	0.77	3.97E-99	395519795	0.77	3.97E-99	395519795
secernin-3 [Myotis lucifugus]	31	0.74	1.00E-95	383087724	0.80	1.07E-103	149022245
secernin-3 [Myotis lucifugus]	32	0.75	7.65E-97	126326616	0.75	5.94E-96	149639530
secernin-3 [Myotis lucifugus]	33	0.74	5.59E-95	327283502	0.57	5.67E-73	62914002
secernin-3 [Myotis lucifugus]	34	0.75	5.94E-96	149639530	0.57	2.08E-72	148695178
secernin-3 [Myotis lucifugus]	35	0.72	9.78E-93	119631548	0.74	5.59E-95	327283502
secernin-3 [Myotis lucifugus]	36	0.74	3.26E-95	224055113	0.63	1.11E-80	301610221
secernin-3 [Myotis lucifugus]	37	0.65	9.04E-83	41054327	0.64	2.14E-82	348519679
secernin-3 [Myotis lucifugus]	38	0.64	1.17E-81	163914461	0.64	1.17E-81	163914461
secernin-3 [Myotis lucifugus]	39	0.63	1.11E-80	301610221	0.02	5.83E+01	358391805
secernin-3 [Myotis lucifugus]	40	0.63	3.32E-80	94482839	0.02	1.08E+02	390350007
secernin-3 [Myotis lucifugus]	41	0.64	2.14E-82	348519679	0.02	5.24E+01	302916981
secernin-3 [Myotis lucifugus]	42	0.62	3.17E-79	61557143	0.03	2.97E+01	340372503
secernin-3 [Myotis lucifugus]	43	0.57	5.67E-73	62914002	0.03	3.34E+01	325145043
secernin-3 [Myotis lucifugus]	44	0.58	5.08E-74	47218098	0.02	1.89E+02	358366518
secernin-3 [Myotis lucifugus]	45	0.57	2.08E-72	148695178	0.03	6.15E+00	326917505
secernin-3 [Myotis lucifugus]	46	0.56	7.73E-72	311272660	0.02	4.93E+01	322710652
secernin-3 [Myotis lucifugus]	47	0.55	8.38E-70	395826584	0.03	8.51E+00	342874004
secernin-3 [Myotis lucifugus]	48	0.55	1.63E-69	327275780	0.02	8.59E+01	147864006
secernin-3 [Myotis lucifugus]	49	0.56	1.52E-70	126308319	0.02	8.02E+01	308469783
secernin-3 [Myotis lucifugus]	50	0.55	6.04E-70	326934081	0.02	6.33E+01	345487083
steroidogenic factor 1 isoform X2 [Camelus ferus]	1	0.86	1.45E-114	395824169	0.95	1.62E-127	325495571
steroidogenic factor 1 isoform X2 [Camelus ferus]	2	0.88	4.90E-117	160221327	0.88	4.90E-117	160221327
steroidogenic factor 1 isoform X2 [Camelus ferus]	3	0.87	5.20E-116	47523442	0.88	2.58E-117	27806027
steroidogenic factor 1 isoform X2 [Camelus ferus]	4	0.87	7.18E-116	344271937	0.87	5.20E-116	47523442
steroidogenic factor 1 isoform X2 [Camelus ferus]	5	0.88	2.58E-117	27806027	0.87	7.18E-116	344271937
steroidogenic factor 1 isoform X2 [Camelus ferus]	6	0.85	4.25E-114	300797824	0.86	1.45E-114	395824169
steroidogenic factor 1 isoform X2 [Camelus ferus]	7	0.95	1.62E-127	325495571	0.85	5.00E-113	332229985
steroidogenic factor 1 isoform X2 [Camelus ferus]	8	0.85	4.96E-114	149047896	0.84	1.46E-112	20070193
steroidogenic factor 1 isoform X2 [Camelus ferus]	9	0.85	1.38E-113	20522231	0.85	4.25E-114	300797824
steroidogenic factor 1 isoform X2 [Camelus ferus]	10	0.85	2.13E-113	74142710	0.84	1.06E-112	297685326
steroidogenic factor 1 isoform X2 [Camelus ferus]	11	0.84	1.82E-112	354499096	0.83	3.52E-111	10945629
steroidogenic factor 1 isoform X2 [Camelus ferus]	12	0.83	3.52E-111	10945629	0.84	9.53E-113	109110256
steroidogenic factor 1 isoform X2 [Camelus ferus]	13	0.84	8.19E-112	1805353	0.84	3.85E-112	216409744
steroidogenic factor 1 isoform X2 [Camelus ferus]	14	0.85	5.00E-113	332229985	0.86	6.84E-115	351702108
steroidogenic factor 1 isoform X2 [Camelus ferus]	15	0.87	7.18E-116	126352395	0.85	6.12E-113	301769265
steroidogenic factor 1 isoform X2 [Camelus ferus]	16	0.84	9.53E-113	109110256	0.84	1.63E-112	384940122
steroidogenic factor 1 isoform X2 [Camelus ferus]	17	0.85	5.85E-113	397473205	0.84	2.79E-112	2077920
steroidogenic factor 1 isoform X2 [Camelus ferus]	18	0.84	1.63E-112	384940122	0.87	7.18E-116	126352395
steroidogenic factor 1 isoform X2 [Camelus ferus]	19	0.84	1.46E-112	20070193	0.85	1.38E-113	20522231

steroidogenic factor 1 isoform X2 [Camelus ferus]	20	0.84	1.06E-112	297685326	0.85	2.13E-113	74142710
steroidogenic factor 1 isoform X2 [Camelus ferus]	21	0.84	1.11E-112	297270159	0.84	8.19E-112	1805353
steroidogenic factor 1 isoform X2 [Camelus ferus]	22	0.84	2.79E-112	2077920	0.76	6.80E-101	355567920
steroidogenic factor 1 isoform X2 [Camelus ferus]	23	0.84	3.85E-112	216409744	0.84	1.82E-112	354499096
steroidogenic factor 1 isoform X2 [Camelus ferus]	24	0.85	6.20E-113	348570102	0.85	6.20E-113	348570102
steroidogenic factor 1 isoform X2 [Camelus ferus]	25	0.85	6.12E-113	301769265	0.85	4.96E-114	149047896
steroidogenic factor 1 isoform X2 [Camelus ferus]	26	0.86	6.84E-115	351702108	0.84	1.11E-112	297270159
steroidogenic factor 1 isoform X2 [Camelus ferus]	27	0.83	1.43E-110	345805854	0.85	5.85E-113	397473205
steroidogenic factor 1 isoform X2 [Camelus ferus]	28	0.76	6.80E-101	355567920	0.83	1.43E-110	345805854
steroidogenic factor 1 isoform X2 [Camelus ferus]	29	0.71	8.06E-94	49036491	0.66	4.93E-87	334311611
steroidogenic factor 1 isoform X2 [Camelus ferus]	30	0.76	4.40E-101	332832881	0.71	8.06E-94	49036491
steroidogenic factor 1 isoform X2 [Camelus ferus]	31	0.66	4.93E-87	334311611	0.76	4.40E-101	332832881
steroidogenic factor 1 isoform X2 [Camelus ferus]	32	0.60	9.24E-79	4586618	0.68	4.11E-90	296190799
steroidogenic factor 1 isoform X2 [Camelus ferus]	33	0.59	9.80E-78	115334528	0.57	2.70E-75	395505691
steroidogenic factor 1 isoform X2 [Camelus ferus]	34	0.59	7.90E-78	45384188	0.59	9.80E-78	115334528
steroidogenic factor 1 isoform X2 [Camelus ferus]	35	0.58	8.89E-76	115529250	0.60	9.24E-79	4586618
steroidogenic factor 1 isoform X2 [Camelus ferus]	36	0.59	5.48E-77	168479587	0.59	7.90E-78	45384188
steroidogenic factor 1 isoform X2 [Camelus ferus]	37	0.56	3.42E-74	4104218	0.59	5.48E-77	168479587
steroidogenic factor 1 isoform X2 [Camelus ferus]	38	0.54	2.97E-71	291565556	0.56	1.15E-73	345326142
steroidogenic factor 1 isoform X2 [Camelus ferus]	39	0.54	5.65E-71	4126870	0.56	3.42E-74	4104218
steroidogenic factor 1 isoform X2 [Camelus ferus]	40	0.56	1.15E-73	345326142	0.58	8.89E-76	115529250
steroidogenic factor 1 isoform X2 [Camelus ferus]	41	0.54	4.09E-71	224809509	0.46	2.48E-59	327290547
steroidogenic factor 1 isoform X2 [Camelus ferus]	42	0.54	5.07E-71	148224522	0.44	1.84E-57	24158439
steroidogenic factor 1 isoform X2 [Camelus ferus]	43	0.68	4.11E-90	296190799	0.45	3.81E-58	66356139
steroidogenic factor 1 isoform X2 [Camelus ferus]	44	0.53	1.95E-69	44355486	0.46	1.83E-59	15145791
steroidogenic factor 1 isoform X2 [Camelus ferus]	45	0.57	2.70E-75	395505691	0.53	1.95E-69	44355486
steroidogenic factor 1 isoform X2 [Camelus ferus]	46	0.68	5.00E-90	149047895	0.47	4.13E-61	1947098
steroidogenic factor 1 isoform X2 [Camelus ferus]	47	0.68	1.25E-89	425578	0.47	2.80E-61	281351212
steroidogenic factor 1 isoform X2 [Camelus ferus]	48	0.68	7.42E-90	148694876	0.43	1.23E-55	218683821
steroidogenic factor 1 isoform X2 [Camelus ferus]	49	0.67	4.13E-89	220401	0.45	1.89E-58	14010847
steroidogenic factor 1 isoform X2 [Camelus ferus]	50	0.48	2.80E-62	350537337	0.46	5.21E-59	13492975

The table shows the top 50 results of BLASTp and RAFTS3 for 5 sequences randomly selected from the test dataset compared against NR. The subject sequences are indicated by their GI number and ordered by the default criteria of each tool; the relative score of each one was calculated.