FASTA Herder: a web application to trim protein sequence sets

Caroline Louis-Jeune[1], Miguel A. Andrade-Navarro[2], and Carol Perez-Iratxeta*[,1,3]

[1]Ottawa Hospital Research Institute, 501 Smyth Road, Ottawa, Ontario K1H 8L6, Canada
[2]Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany
[3]Department of Cellular and Molecular Medicine, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
*Corresponding author's e-mail address: cpereziratxeta@gmail.com

Published online: April 28, 2014 (version 1)
Cite as: Louis-Jeune et al., ScienceOpen Research 2014 (DOI: 10.14293/A2199-1006.01.SOR-LIFE.67837.v1)
Reviewing status: Please note that this article is under continuous review. For the current reviewing status and the latest referee's comments please click **here** or scan the QR code at the end of this article.

Primary discipline: Life Sciences
Associated discipline: Quantitative & Systems biology, Bioinformatics & Computational biology
Keywords: Multiple sequence alignment, BLAST, FASTA, Software, Web service, Sequence similarity

ABSTRACT

The ever increasing number of sequences in protein databases usually turns out large numbers of homologs in sequence similarity searches. While information from homology can be very useful for functional prediction based on amino acid conservation, many of these homologs usually have high levels of identity among themselves, which hinders multiple sequence alignment computation and, especially, visualization. More generally, high redundancy reduces the usability of a protein set in machine learning applications and biases statistical analyses. We developed an algorithm to identify redundant sequence homologs that can be culled producing a streamlined FASTA file. As a difference from other automatic approaches that only aggregate sequences with high identity, our method clusters near full-length homologs allowing for lower sequence identity thresholds. Our method was fully tested and implemented in a web application called FASTA Herder, publicly available at http://fh.ogic.ca.

INTRODUCTION

Multiple sequence alignment (MSA) remains the most important analytic tool to assess evolutionary relations between proteins and to determine the conserved regions of the sequence that usually harbor structural and functional properties. However, the algorithms that produce MSAs are not perfect and further manual curation carried through by visual inspection is necessary [1]. Inspection is also required to detect or confirm some features of the aligned proteins, such as the presence of a structural domain. Notwithstanding the availability of excellent MSA visualization tools [2], the growth of the protein database is making difficult these tasks, as well as the calculation of the alignment itself (e.g. Uniref100 [3], which currently contains over 23 million sequences, has nearly doubled every two years since 2008). Removing highly redundant sequences in a set of homologs helps MSA and its interpretation as well as reducing biases in the protein set when it is taken as a sample for statistical analysis or for machine learning applications. In a previous work we conceived an algorithm to cluster the whole protein database by aggregating near full-length homologs [4]. Instead of clustering sequences based just on high sequence identity, our approach allows low but significant levels of identity according to BLAST [5], and is very restrictive in requiring this similarity to be full length without large unmatched portions between the compared sequences. This is to ensure that when two sequences are aggregated they have the same domain structure, which increases the likelihood of both having a similar function [6]. We have implemented this algorithm in a web tool to quickly organize and/or trim a set of proteins in FASTA format.

APPROACH

Given a set of proteins, we use a greedy algorithm to group similar sequences together. First, we pick the longest sequence as the query sequence, and we compare it with BLAST to the subset of sequences that either have the same length as the query or are shorter by less than n amino acids (aa). We will retain BLAST matches covering the full length of the query or shorter by less than $n/2$ aa from each end. The value of n depends on the length of the query sequence, q, and is by default equal to 32, if $q \geq 200$; 20, if $200 > q \geq 100$; 10, if $100 > q \geq 60$; 7, if $60 > q \geq 40$; 5, if $40 > q \geq 20$; and 2, if $q < 20$. Sequences are considered to match the query sequence if they are sufficiently similar to the query as detected by a BLAST single hit (either continuous or gapped) that covers all (or almost all) of the protein. An identity threshold of 24.8% was used for BLAST hits if $q \geq 80$, else a function of q equal to $290.15 \cdot q^{-0.562}$ for shorter hits, following Sander and Schneider [7]. Query and matched sequences are aggregated and removed from the set and the longest sequence among the remaining ones is chosen as a new query. This procedure is repeated until no sequences remain.

BENCHMARKS

To validate the reduction in the number of sequences in FASTA files using our method, we used the OrthoBench, a published benchmark set of manually curated protein families that was designed to assess orthologs' detection methods [8] and is available at http://eggnog.embl.de/orthobench/. The OrthoBench consists of 1677 proteins from 12 Metazoa species grouped in 70 protein families classified according to the rate of evolution, domain architecture, low complexity/ repeats, lineage-specific loss/duplication, and alignment quality (Supplementary Table S1). We seeked to evaluate how our algorithm avoids clustering together sequences from different families despite allowing low levels of homology. Hence, we noted the level of reduction in the number of sequences and the number of errors (when sequences belonging to different families are clustered together) as performance indicators. The overall method performance was good, reducing the number of sequences down to 35.24% of that in the original FASTA size with 0 misclassifications with the default parameters (Supplementary Table S2). Some proteins contain low-complexity regions (LCRs), regions with repetitive sequences and little diversity in their amino acid composition. Hypothesizing that the length of LCRs will be extremely variable within protein families and might cause misclassification of orthologs in different clusters, we tested how detecting and masking these from the original sequences would affect the results. For conservative parameters of LCR detection, our algorithm achieved a similar reduction in the number of sequences without increasing the number of errors (see supplement material and Supplementary Table S3). As this is heavily dependent on particular protein sets, we decided to provide LCR detection and filtering as a user option. Finally we tested how different thresholds of tolerance of length differences between sequences allowed to cluster together would affect performance. Results indicate that moderately increasing tolerance increases the reduction in the number of sequences without largely affecting the number of misclassifications (i.e. reduction of 30.47% with 1 misclassification; see Supplementary Table S4). The number of misclassifications is only indicative because in the OrthoBench, proteins from different families can be highly divergent, and therefore the number of mistakes could be expected to be smaller than in real-life applications.

WEB TOOL

FASTA Herder accepts a protein set in FASTA file format and quickly clusters it (e.g. the OrthoBench is clustered in less than 50 seconds). Optionally, LCRs, as detected with SEG [9], can be ignored before clustering. The server also permits to adjust the stringency of the clustering based on the allowed difference in length between sequences to be clustered together, although it may be important to stick to conservative levels when culling sequences from an MSA. We use the BLAST implementation of the BLAST+ suite [10].

DISCUSSION

We have implemented an ease-of-use web application to quickly reduce the redundancy of a protein set by clustering homologs of comparable lengths. Although our work is somewhat related to the field of orthology prediction methods (discussed in Ref. [8]), its purpose and scope are different. Orthology detection encompasses the use of whole genomes to identify proteins derived from a single ancestral sequence through speciation events. The aim of our tool is to reduce potentially high redundancy in a FASTA file that may contain proteins belonging to one or more families. This would speed MSA calculation, facilitate MSA inspection, and remove biases that may affect any statistical analysis or computational application involving that set of proteins. A previously published web tool that possess a capability similar to this work is PISCES [11], a protein database that culls the Protein Data Bank to build the largest possible set of structures that comply with identity cut-offs for sequence and structure. More related to our tool, PISCES accepts a FASTA file as well and culls out sequences by a single identity threshold that can be provided by the user. We compared the reduction in the number of sequences by PISCES and FASTA Herder on the OrthoBench set with the default parameters. We had to split the OrthoBench set in four parts because it cannot be handled by PISCES. With the default parameters, FASTA Herder reduced much less the number of sequences in every part. However, PISCES did often misclassify sequences and was much slower (see Supplementary Table S5). In conclusion, we have created a tool to remove redundant sequences from sets of sequences. Simplification of MSAs was our main goal, motivated by the increase in redundant sequences in the databases due to genomic sequencing projects. Our method is simple, works in a matter of seconds with large datasets that are unmanageable by other methods, and includes a number of options (LCR filtering, length thresholds) that make it very flexible.

SUPPLEMENTARY INFORMATION

Supplementary material is available **here**.

ACKNOWLEDGMENTS

This work was supported (in part) by a Government of Ontario Ministry of Research and Innovation grant (ORF-RE05-084).

REFERENCES

[1] Thompson JD, Linard B, Lecompte O, Poch O. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS ONE. 2011;6: e18093. doi:10.1371/journal.pone.0018093.t001

[2] Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, Barton GJ. Visualization of multiple alignments, phylogenies and gene family evolution. Nat Methods. 2010;7:s16–s25. doi:10.1038/nmeth.1434

[3] Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters.

Bioinformatics. 2007;23:1282–8. doi:10.1093/bioinformatics/btm098

[4] Perez-Iratxeta C, Palidwor G, Andrade-Navarro MA. Towards completion of the Earth's proteome. EMBO Rep. 2007;8:1135–41. doi:10.1038/sj.embor.7401117

[5] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402. doi:10.1093/nar/25.17.3389

[6] Ponting CP, Schultz J, Copley RR, Andrade MA, Bork P. Evolution of domain families. Adv Protein Chem. 2000;54:185–244. doi:10.1016/S0065-3233(00)54007-8

[7] Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 1991;9:56–68. doi:10.1002/prot.340090107

[8] Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, Bork P. Orthology prediction methods: a quality assessment using curated protein families. Bioassays. 2011;33:769–80. doi:10.1002/bies.201100062

[9] Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Meth Enzymol. 1996;266:554–71.

[10] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. doi:10.1186/1471-2105-10-421

[11] Wang G, Dunbrack RL. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res. 2005;33:W94–W98. doi:10.1093/nar/gki402

COMPETING INTERESTS

The authors declare no competing interests.

PUBLISHING NOTES

Please note that this article may not have been peer reviewed yet and is under continuous post-publication peer review. For the current reviewing status please click **here** or scan the QR code on the right.