BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 15 2009, pages 1972–1973
doi:10.1093/bioinformatics/btp348

Phylogenetics

trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses

Salvador Capella-Gutiérrez, José M. Silla-Martínez and Toni Gabaldón*

Comparative Genomics group, Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Dr. Aiguader, 88 08003 Barcelona, Spain

Received on May 27, 2008; revised on May 20, 2009; accepted on June 1, 2009

Advance Access publication June 8, 2009

Associate Editor: Martin Bishop

ABSTRACT

Summary: Multiple sequence alignments are central to many areas of bioinformatics. It has been shown that the removal of poorly aligned regions from an alignment increases the quality of subsequent analyses. Such an alignment trimming phase is complicated in large-scale phylogenetic analyses that deal with thousands of alignments. Here, we present trimAl, a tool for automated alignment trimming, which is especially suited for large-scale phylogenetic analyses. trimAl can consider several parameters, alone or in multiple combinations, for selecting the most reliable positions in the alignment. These include the proportion of sequences with a gap, the level of amino acid similarity and, if several alignments for the same set of sequences are provided, the level of consistency across different alignments. Moreover, trimAl can automatically select the parameters to be used in each specific alignment so that the signal-to-noise ratio is optimized.

Availability: trimAl has been written in C++, it is portable to all platforms. trimAl is freely available for download (http://trimal.cgenomics.org) and can be used online through the Phylemon web server (http://phylemon2.bioinfo.cipf.es/). Supplementary Material is available at http://trimal.cgenomics.org/publications.

Contact: tgabaldon@crg.es

Multiple sequence alignments (MSA) are central to many areas of bioinformatics, including phylogenetics, homology modeling, database searches and motif finding. Recently, such MSA-based techniques have been incorporated in high-throughput pipelines such as genome annotation and phylogenomics analyses. In all these applications, the reliability and accuracy of the analyses depend critically on the quality of the underlying alignments. A plethora of computer programs and algorithms for MSA are currently available (Notredame, 2007), which implement different heuristics to find mathematically optimal solutions to the MSA problem. Accuracies of 80–90% have been reported for the best algorithms, but even the best scoring alignment algorithms may fail with certain protein families or at specific regions in the alignment. The situation worsens in large-scale analyses, where faster but less reliable algorithms and large numbers of automatically selected sequences are used. It is therefore generally assumed that trimming the alignment, so that poorly aligned regions are eliminated, increases the accuracy of the resulting MSA-based applications (Talavera and Castresana, 2007). Some programs such as G-blocks (Castresana, 2000) have been developed to assist in the MSA trimming phase by selecting blocks of conserved regions. They have become very popular and are extensively used, with good performance, in small-to-medium scale datasets, where several parameters can be tested manually (Talavera and Castresana, 2007). However, their use over larger datasets is hampered by the need for defining, prior to the analysis, the set of parameters that will be used for all sequence families. Here, we present trimAl, a tool for automated alignment trimming. Its speed and the possibility for automatically adjusting the parameters to improve the phylogenetic signal-to-noise ratio, makes trimAl especially suited for large-scale phylogenomic analyses, involving thousands of large alignments.

trimAl has been developed in a GNU/Linux environment using C++ programming language and has been tested on various UNIX, Mac and Windows platforms. Moreover, we have developed a web server to run trimAl online (http://phylemon2.bioinfo.cipf.es/), which has been included in the Phylemon suite for phylogenetic and phylogenomic tools (Tarraga *et al*., 2007). The documentation, source files and additional information for trimAl are available through a wiki page (http://trimal.cgenomics.org).

trimAl reads and renders protein or nucleotide alignments in several standard formats. trimAl starts by reading all columns in an alignment and computes a score (S_x) for each of them. This score can be a gap score (S_g), a similarity score (S_s) or a consistency score (S_c). The score for each column can be computed based only on the information from that column or, if a window size of w is specified, it corresponds to the average value of w columns around the position considered.

The gap score (S_g) for a column is the fraction of sequences without a gap in that position. The residue similarity score (S_s) consists of mean distance (MD) scores as described in Thompson *et al*. (2001) and Supplementary Material. This score uses the MD between pairs of residues, as defined by a given scoring matrix. Finally, the consistency score (S_c) can only be computed when more than one alignment for the same set of sequences is provided. Details on how these scores are computed are provided in the Supplementary Material. In brief, S_c measures the level of consistency of all the residue pairs found in a column as compared with the other alignments. The alignment with the highest consistency is chosen and then trimmed to remove the columns that are less conserved, according to S_c or other thresholds set by the user.

*To whom correspondence should be addressed.

Once all column scores have been computed trimAl can proceed in two ways. If both a score and a minimum conservation threshold are provided, trimAl renders a trimmed alignment in which only the columns with scores above the *score threshold* are included, as far as the number of selected columns is above a *conservation threshold* defined by the user. If this number is below the *conservation threshold*, trimAl will add more columns to the trimmed alignment in a decreasing order of scores until the *conservation threshold* is reached. The *conservation threshold* corresponds to the minimum percentage of columns, from the original alignment, which the user wants to include in the trimmed alignment.

Alternatively, if the automatic selection of parameters options is selected, trimAl will compute specific score thresholds depending on the inherent characteristics of each alignment. So far, trimAl incorporates three modes for the automated selection of parameters, *gappyout*, *strict* and *strictplus*, which are based on the different use of gap and similarity scores. Moreover, the option *automated1* implements a heuristic to decide the most appropriate mode depending on the alignment characteristics. The heuristics to define such parameters have been designed based on the results of a benchmark. Details on the heuristics and the benchmark can be found in the online documentation of the program. In brief, the automatic selection of parameters approximate optimal cutoffs by plotting, internally, the cumulative graphs of gap and similarity scores of the columns in the alignment (see online documentation).

We expanded, using ROSE simulations (Stoye *et al*., 1998) a benchmark set that has been used previously to test the improvement in phylogenetic performance after an alignment trimming phase (Talavera and Castresana, 2007). This dataset simulates several evolutionary scenarios varying in the number and length of the sequences, the topology of the underlying tree and the level of sequence divergence considered. We compared the results obtained from MUSCLE alignments before and after trimming with trimAl using automated selection of parameters. The accuracy of the resulting trees was measured by comparing them with the original trees used to generate the sequence sets, and measuring the Robinson Foulds distance (Robinson and Foulds, 1981). We observed an overall improvement of the phylogenetic accuracy after trimming. Using *-automated1* option of trimAl, the trimmed alignment always produced Maximum Likelihood trees that were of equal (36%) or significantly better (64%) quality as compared with the tree derived from the complete alignment. For Neighbor Joining reconstruction the *-strictplus* option of trimAl worked best, improving the phylogenetic accuracy in 89% of the scenarios. In most scenarios (90%), trimAl outperformed Gblocks v0.91b with default parameters. Most importantly, the use of Gblocks default parameters diminished the accuracy of the subsequent tree reconstruction in half of the scenarios considered. In contrast, the use of trimAl automated methods rarely (1.5%) undermined the topological accuracy of the resulting phylogenetic tree (see Supplementary Material for more details).

To test the applicability of trimAl on real datasets as well as its suitability for large-scale phylogenetic datasets, we ran trimAl on the complete set of MUSCLE alignments generated for the Human Phylome project (Huerta-Cepas *et al*., 2007). This includes a total of 31 182 alignments, containing, on average, 67 sequences of 1472 positions of length. Trimming these alignments using the *-gappyout* and *automated1* options used 5 min 45 s and 125 min, 2 s, respectively, on a computer with an Intel QuadCore XEON E5410 processors and 8 GB of RAM.

trimAl has been used previously in a pipeline to reconstruct complete collections of gene trees. In this case, the parameter sets used were a minimum conservation threshold of 60% and a gap threshold of 90% (-cons 60 -gt 0.9). Complete and trimmed alignments used to generate the phylomes included in PhylomeDB (Huerta-Cepas *et al*., 2008) can be viewed through this database.

ACKNOWLEDGEMENTS

We acknowledge Juan M. Garcia-Gomez and members of the Gabaldón group for fruitful discussions and suggestions. The authors also thank Jordi Burguet-Castell, Pablo Escobar and Joaquín Tàrraga for their technical assistance and José Castresana for providing the data necessary for the benchmark.

Funding: FIS (06-213 to T.G.) and MEC (GEN2006-27784-E/PAT to T.G.).

Conflict of Interest: none declared.

REFERENCES

Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.

Huerta-Cepas,J. *et al.* (2007) The human phylome.*Genome Biol.*, **8**, R109.

Huerta-Cepas,J. *et al.* (2008) PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Res.*, **36**, D491–D496.

Notredame,C. (2007) Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput. Biol.*, **3**, e123.

Robinson,D.R. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

Stoye,J. *et al.* (1998) Rose: generating sequence families. *Bioinformatics*, **14**, 157–163.

Talavera,G. and Castresana,J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, **56**, 564–577.

Tarraga,J. *et al.* (2007) Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics. *Nucleic Acids Res.*, **35**, W38–W42.

Thompson,J.D. *et al.* (2001) Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.,* **314**, 937–951.