MaxAlign: maximizing usable data in an alignment

BACKGROUND: The presence of gaps in an alignment of nucleotide or protein sequences is often an inconvenience for bioinformatical studies. In phylogenetic and other analyses, for instance, gapped columns are often discarded entirely from the alignment. RESULTS: MaxAlign is a program that optimizes the alignment prior to such analyses. Specifically, it maximizes the number of nucleotide (or amino acid) symbols that are present in gap-free columns - the alignment area - by selecting the optimal subset of sequences to exclude from the alignment. MaxAlign can be used prior to phylogenetic and bioinformatical analyses as well as in other situations where this form of alignment improvement is useful. In this work we test MaxAlign's performance in these tasks and compare the accuracy of phylogenetic estimates including and excluding gapped columns from the analysis, with and without processing with MaxAlign. In this paper we also introduce a new simple measure of tree similarity, Normalized Symmetric Similarity (NSS) that we consider useful for comparing tree topologies. CONCLUSION: We demonstrate how MaxAlign is helpful in detecting misaligned or defective sequences without requiring manual inspection. We also show that it is not advisable to exclude gapped columns from phylogenetic analyses unless MaxAlign is used first. Finally, we find that the sequences removed by MaxAlign from an alignment tend to be those that would otherwise be associated with low phylogenetic accuracy, and that the presence of gaps in any given sequence does not seem to disturb the phylogenetic estimates of other sequences. The MaxAlign web-server is freely available online at http://www.cbs.dtu.dk/services/MaxAlign where supplementary information can also be found. The program is also freely available as a Perl stand-alone package.

General information
State: Published
Organisations: Center for Biological Sequence Analysis, Department of Systems Biology
Contributors: Oliveira, R. G., Sackett, P. W., Pedersen, A. G.
Pages: 312
Publication date: 2007
Peer-reviewed: Yes

Publication information
Journal: BMC Bioinformatics
Volume: 8
ISSN (Print): 1471-2105
Ratings:
BFI (2019): BFI-level 1
Web of Science (2019): Indexed yes
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Scopus rating (2017): CiteScore 2.49 SJR 1.479 SNIP 0.878
Web of Science (2017): Impact factor 2.213
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 1
Scopus rating (2016): CiteScore 2.54 SJR 1.581 SNIP 0.974
Web of Science (2016): Impact factor 2.448
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): CiteScore 2.77 SJR 1.737 SNIP 1.079
Web of Science (2015): Impact factor 2.435
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 1
Scopus rating (2014): CiteScore 2.91 SJR 1.916 SNIP 1.185
Web of Science (2014): Impact factor 2.576
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): CiteScore 3.38 SJR 1.999 SNIP 1.323
Web of Science (2013): Impact factor 2.672
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): CiteScore 3.24 SJR 1.9 SNIP 1.145