molecular systems biology

Manuscript EMBO-2011-3104

Fast, scalable generation of high quality protein multiple sequence alignments using Clustal Omega

Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, Desmond G Higgins

Corresponding author: Desmond Higgins, University College Dublin

Review timeline:

Submission date:	23 July 2011
Editorial Decision:	14 August 2011
Revision received:	23 August 2011
Accepted:	6 September 2011

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 14 August 2011

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the three referees whom we asked to evaluate your manuscript. As you will see from the reports below, the referees appreciate the performance of clustal omega when performing alignments of very large number of sequences and the EPA functionality. They note that the software appears to match the increasing demands created by current ongoing large sequencing efforts and reviewer #2 feels that it is likely to become a standard tool similar to the previous clustal versions. On the other hand, the manuscript remains rather technical and represents the implementation of a methodological concept that has been presented before. As such, I am afraid we could not offer to publish the manuscript as a full research article. However, in view of the balance of opinions expressed by the reviewers, we feel that we can consider a shortened version of the manuscript if reformatted as a short Report (approx. 25000 characters, 3 figures maximum, <http://www.nature.com/msb/authors>). Here would be some possible recommendations that may help you to shorten the text and make it more compact:

-the introduction could be considerably shortened for example by reducing/omitting the extensive historical account of MSA methods
- the Results and Discussion sections can be combined
-the results would benefit from being streamlined to focus on key performance results and functionalities of clustal omega.
-While Materials and Method section is not included in the character count, we feel that this section should also be written in a much more compact manner. The list of benchmark is useful but their detailed description would be better suited for Supplementary information

*** PLEASE NOTE *** As part of the EMBO Publications transparent editorial process initiative (see our Editorial at http://www.nature.com/msb/journal/v6/n1/full/msb201072.html), Molecular Systems Biology will publish online a Review Process File to accompany accepted manuscripts. When preparing your letter of response, please be aware that in the event of acceptance, your cover letter/point-by-point document will be included as part of this File, which will be available to the scientific community. More information about this initiative is available in our Instructions to Authors. If you have any questions about this initiative, please contact the editorial office msb@embo.org.

If you feel you can satisfactorily deal with these points and those listed by the referees, you may wish to submit a revised version of your manuscript. Please attach a covering letter giving details of the way in which you have handled each of the points raised by the referees.

Thank you for the opportunity to examine this work and I look forward to receiving your revised manuscript.

Yours sincerely,

Editor
Molecular Systems Biology

REFEREE REPORTS

Reviewer #1 (Remarks to the Author):

Report for Sievers et al. MSB 2011

In this study the authors present a new algorithm for multiple sequence alignment. While MSA has been an active research area for a while, the authors present a method that allows for the alignment of very large numbers of sequences, in the 100,000s as may be required by ongoing sequencing efforts.

I should start by pointing out that this strikes me as a very technical paper, not of the sort that I would usually expect to see in MSB, though this is certainly an editorial and not a refereeing decision. Moreover, please note that I am an "educated consumer" of MSA programs, i.e., I have some high-level understanding of how the different algorithms work and have extensively use many of them, but am not familiar with any technical details. Hence I review this paper from a "consumer" perspective.

This algorithm strikes me as beautiful, and the paper is written very clearly, enabling even me to appreciate its elegance. Having said that, the elegant idea (of embedding sequences in a ~log(N) dimensional space) used here appears to have been presented in a previous paper from the same laboratory.

Most of the paper is spent on reporting benchmarking (in terms of both runtime and accuracy) of the algorithm as compared to other methods. The overall message appears to be that ClustalOmega appears to be the fastest and most accurate algorithm for very large alignments, while it is still substantially outperformed by other methods for smaller alignments. It also offers the Profile Alignment use-case, which is likely to be more important in the future.

Looking at the results presented here (again, from the standpoint of a "consumer"), I would not personally expect to make much use of this algorithm and will probably continue to use MAFFT or MUSCLE, mainly since they're still performing similarly well or better for smaller alignments, which are the main use for myself. I expect this to be true for the majority of users. However, this may change in the future and especially for those who work on major sequencing / data repository centers (such as the EBI), as sequencing of more species ramps up.

In summary, the authors present a novel and beautiful algorithm that may be of limited added value now, but will likely have stronger impact in the further future.

Reviewer #2 (Remarks to the Author):

This paper introduces Clustal Omega, a new implementation of the widely-used Clulstal software package for multiple sequence alignment (MSA). The main novelty of this program compared to previous versions of Clustal is its ability to deal with huge data sets. This is mainly achieved by using an efficient clustering algorithm to efficiently calculate a guide tree for progressive alignment. This method has been published previously by the same group.

The work described in this manuscript is an important step in the development of MSA algorithms. With the huge datasets that are now available, life sciensits are in urgent need of MSA programs for large sequence sets. As the authors explain, constructing the guide tree for progressive alignment is the bottleneck of most MSA programs: calculating a tree from the input sequences with traditional methods takes (at least) O(n^2) time while progressively aligning the sequences according to this tree can be done in linear time.

A second interesting novelty of Clustal Omega is the option to use previously calculated "External Profile Alignments" (EPA) to improve the alignment procedure. This is done using an HMM approach developed by one of the co-authors. Since for many protein families, there are now profile HMMs available, it is a very sensible idea to use this information for improved MSA, rather than relying on sequence similarity alone, as do traditional alignment methods.

Some of the important conceptual novelties in Clustal Omega, compared to previous versions of Clustal, have been previously published, such as the mBed clustering approach. Nevertheless, the implementation of Clustal Omega is, of course, new. I think this is a very important step in the development of MSA methods and the described tool will certainly become a standard method for life scientists, as previous versions of Clustal before.

The paper is well written and clear and the test examples convincingly demonstrate the efficiency of the developed method. Thus, I strongly support publication of this manuscript.

1st Revision - authors' response 23 August 2011

We wish to thank you for dealing with this so quickly. Basically, you wanted us to shorten the ms. to a Brief Report. We have now done this and hope this has been done appropriately. In summary, we shortened the text to <25000 characters and have 3 figures:

1) moved most of the methods section and 1 figure to a supplementary material section

2) merged 2 old figures into one thus giving three figures in the main paper (plus 3 in supplemental; one figure is new)

3) reduced the text to <25,000 characters. This includes cover page, references and figure legends but excludes tables and methods.

