BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 2 2010, pages 263–265
doi:10.1093/bioinformatics/btp651

Sequence analysis

AQUA: automated quality improvement for multiple sequence alignments

Jean Muller[1,*,†,‡], Christopher J. Creevey[1,†,§], Julie D. Thompson[2], Detlev Arendt[1] and Peer Bork[1,3]

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany, [2]Département de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, (CNRS/INSERM/ULP), BP 10142, 67404 Illkirch Cedex and [3]Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Strasse 10, 13092 Berlin, Germany

Received on August 11, 2009; revised on October 21, 2009; accepted on November 15, 2009

Advance Access publication November 19, 2009

Associate Editor: John Quackenbush

ABSTRACT

Summary: Multiple sequence alignment (MSA) is a central tool in most modern biology studies. However, despite generations of valuable tools, human experts are still able to improve automatically generated MSAs. In an effort to automatically identify the most reliable MSA for a given protein family, we propose a very simple protocol, named AQUA for 'Automated quality improvement for multiple sequence alignments'. Our current implementation relies on two alignment programs (MUSCLE and MAFFT), one refinement program (RASCAL) and one assessment program (NORMD), but other programs could be incorporated at any of the three steps.

Availability: AQUA is implemented in Tcl/Tk and runs in command line on all platforms. The source code is available under the GNU GPL license. Source code, README and Supplementary data are available at http://www.bork.embl.de/Docu/AQUA.

Contact: muller@embl.de, bork@embl.de

1 INTRODUCTION

Multiple sequence alignment (MSA) of protein sequences is a central tool in most modern biology studies. However, despite generations of valuable tools, so far there has been no optimal solution and human experts [using external information such as 3D structures (O'Sullivan *et al*., 2004)] are still able to improve automatically generated MSAs (Pirovano *et al*., 2008; Waterhouse *et al*., 2009). The next-generation sequencing technologies are now producing massive amounts of data so that for large-scale analysis, technical issues such as speed also becomes important parameters. High-throughput comparative analyses require automated and fast pipelines that include numerous MSAs as a starting point for structural and functional studies (Thompson and Poch, 2006) and phylogenomic approaches (Dunn *et al*., 2008). Automatic high-quality MSAs are crucial to guarantee the reliability and success of such studies. Numerous methods are available to build MSAs [e.g. MUSCLE (Edgar, 2004), MAFFT (Katoh and Toh, 2008), ProbCons (Do *et al*., 2005) and PRANK (Loytynoja and Goldman, 2008)], to refine them [e.g. RASCAL (Thompson *et al*., 2003) and REFINER (Chakrabarti *et al*., 2006)] and to assess their quality [e.g. NORMD (Thompson *et al*., 2001) and MUMSA (Lassmann and Sonnhammer, 2005)]. Nevertheless, despite the fact that these methods often produce reliable results, some programs appear to be more suitable for particular situations than others (Thompson *et al*., 2005). In an effort to automatically identify the most reliable MSA for a given protein family, we propose AQUA (Automated QUality improvement for multiple sequence Alignments) a very simple protocol, combining existing tools in the following three steps:

- Computation of an initial MSA
- Refinement of the MSA
- Evaluation and selection of the best MSA

AQUA has been evaluated using carefully validated sets of MSAs and applied to a large-scale dataset.

2 METHODS AND DATASETS

AQUA is described as follows (Fig. 1a): given a set of protein sequences, we (i) first align them using one (or more) alignment program(s). In our implementation, we use Muscle (v3.7) and Mafft (v6.611). Both of these programs make use of heuristics to perform fast and reliable MSA, which nevertheless can contain some errors. Therefore, (ii) a refinement step is introduced to detect and correct these errors. For this post-processing, we use the Rascal (v1.34) program. Finally, (iii) the NORMD (v1.3) program is used to estimate the quality of each MSA produced in the previous steps. The NorMD score gives information about the general quality of the alignment [a NorMD >0.6 indicates a reliable MSA (Thompson *et al*., 2003)], but can also be used to compare different versions of the same MSA. Thus, we select the MSA with the highest NorMD value. We used the default parameters for each program. Nevertheless, for large sets of alignments, a mixture of different algorithms, parameter settings and post-processing

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

‡Département de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, (CNRS/INSERM/ULP) and Laboratoire de Diagnostic Génétique, CHU Strasbourg Nouvel Hôpital Civil, 1 place de l'hôpital, 67000 Strasbourg, France.

§Teagasc Animal Bioscience Centre, Grange, Dunsany, Co. Meath, Ireland.

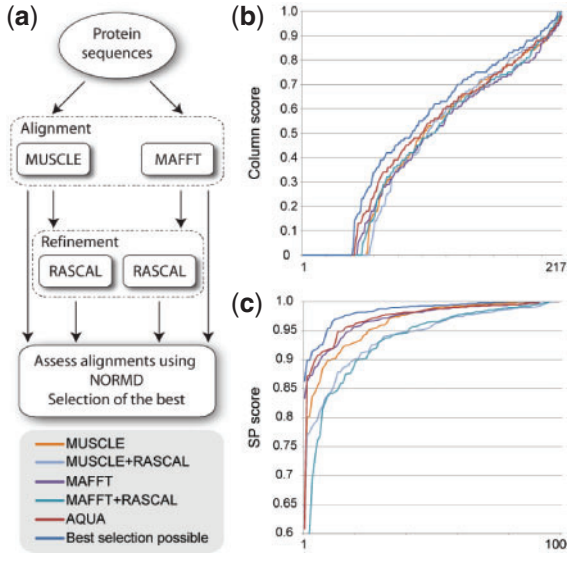

Fig. 1. (**a**) Flowchart showing the strategy used in AQUA. 'AQUA' corresponds to the alignment quality obtained using our strategy (i.e selecting the program with the highest NorMD). 'Best selection possible' corresponds to the highest possible alignment quality (i.e selecting the program with the highest score depending on the dataset used). (**b**) Quality distribution plot using BAliBASE 3.0 based on column scores. (**c**) Quality distribution plot for single copy orthologs based on SP scores. Scores are plotted for each method in ascending order.

operations could be implemented. AQUA was tested using the following datasets (see Supplementary Information):

- Dataset 1 is BAliBASE 3.0 (Thompson *et al*., 2005), a well-established benchmark with 6255 proteins in 217 MSAs covering the major problems encountered when building MSA. The MSAs computed by AQUA were compared with the reference alignment provided, using the column score (the percentage of correctly aligned columns in the MSA), which tests the ability of the programs to align all the sequences correctly and thus provides a greater distinction between the programs (Thompson *et al*., 1999).
- Dataset 2 contains 100 alignments from single copy orthologs (1 protein per genome) of 19 Metazoa amended with putative orthologs from EST data of seven further animals (Creevey *et al*., manuscript in preparation), i.e we added frequently incomplete sequences to increase the demands on the dataset. The orthologs were taken from the eggNOG database and belong to a subset of metazoan non-supervised orthologous groups (meNOGs) (Jensen *et al*., 2008). Reference alignments for this set were constructed manually and the accuracy of our automatic alignments was assessed by comparing them with these references. Here, the many sequence fragments lead to MSAs with few complete columns, and hence the column score is unsuitable. Therefore, we use the sum-of-pairs (SP) score (the percentage of correctly aligned residue pairs in the MSA) (Thompson *et al*., 1999).
- Dataset 3 corresponds to the full meNOG dataset comprising 20 262 MSAs connecting 247 040 proteins from 19 metazoan. The dataset is used to evaluate the performance and the contribution of the different programs.

3 RESULTS AND DISCUSSION

The column score distributions of the BAliBASE set of alignments (Fig. 1b) confirm that AQUA is able to increase the global MSA quality compared with individual programs. As expected, AQUA has the greatest effect on the more divergent cases (i.e. lower column scores) and the smallest effects in highly conserved regions.

When analyzing the 100 alignments from metazoan single copy orthologs, AQUA selects 62 MSAs from Mafft, 11 from Mafft+Rascal, 18 from Muscle and 9 from Muscle+Rascal. Figure 1c compares the resulting alignments with 100 manually constructed MSAs. The SP score distributions clearly show that AQUA provides alignment quality closer to what a skilled human being would create than any of the programs individually.

Finally, the full meNOG dataset was used to assess AQUA in high throughput. We noticed that >98% of the MSA computed have a NorMD score >0.6, emphasizing the overall high quality of the dataset. The MSAs selected are as follows: 9317 Muscle, 1935 Muscle+Rascal, 8011 Mafft and 999 Mafft+Rascal. The results imply first that the refinement step can increase the quality of the initial MSA in ∼15% of the cases and second that the employment of more than one alignment program will considerably improve the overall quality of large sets of MSAs.

Taken together, we showed that AQUA represents a simple and reliable method to obtain a high-quality MSA. It can be applied to a single alignment program with little extra computational cost or to a selection of complementary alignment programs [e.g. 2D aware programs such as PRALINE (Simossis and Heringa, 2005) or Expresso (Armougom *et al*., 2006)] in parallel to ensure the best quality final alignment, at the cost of increased computer time. Future work will include the optimization of alignment programs and parameters, as well as the construction of a new MSA by combining optimally aligned regions from different candidate MSA. In any case, AQUA not only gives an indication of the quality of MSAs in large datasets, but their automatic improvement should increase the signal-to-noise ratio in many genomics, proteomics and phylogenetic studies.

Funding: Bundesministerium für Bildung und Forschung QuantPro (grant 0313831D).

Conflict of Interest: none declared.

REFERENCES

Armougom,F. *et al*. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res*., **34**, W604–W608.

Chakrabarti,S. *et al*. (2006) Refining multiple sequence alignments with conserved core regions. *Nucleic Acids Res*., **34**, 2598–2606.

Do,C.B. *et al*. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res*., **15**, 330–340.

Dunn,C.W. *et al*. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745–749.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.

Jensen,L.J. *et al*. (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*., **36**, D250–D254.

Katoh,K. *et al*. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.*, **9**, 286–298.

Lassmann,T. and Sonnhammer,E.L.L. (2005) Automatic assessment of alignment quality. *Nucleic Acids Res*., **33**, 7120–7128.

Loytynoja,A. *et al*. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–1635.

O'Sullivan,O. *et al*. (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**, 385–395.

Pirovano,W. *et al*. (2008) The meaning of alignment: lessons from structural diversity. *BMC Bioinformatics*, **9**, 556.

Simossis,V.A. and Heringa,J. (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res*., **33**, W289–W294.

Thompson,J.D. *et al*. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*., **27**, 2682–2690.

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

Thompson,J.D. *et al*. (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*, **19**, 1155–1161.

Thompson,J.D. *et al*. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.

Thompson,J.D. and Poch,O. (2006) Multiple sequence alignment as a workbench for molecular systems biology. *Curr. Bioinform.*, **1**, 95–104.

Waterhouse,A.M. *et al.* (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.