BIOINFORMATICS APPLICATIONS NOTE Vol. 15 no. 10 1999
Pages 862-863

MULTICLUSTAL: a systematic method for surveying Clustal W alignment parameters

Jeffrey Yuan, Angela Amend, Joseph Borkowski, Reynold DeMarco, Wendy Bailey, Yuan Liu, Guochun Xie and Richard Blevins

Merck and Co., Inc., Department of Bioinformatics, P.O. Box 2000 - RY80-A1, Rahway, NJ 07065, USA

Received on February 11, 1999; revised on April 29, 1999; accepted on May 10, 1999

Abstract

Summary: *MULTICLUSTAL is a Perl script designed to automate the process of alignment parameter choice for Clustal W with the goal of generating high quality multiple sequence alignments.*

Availability: *Freely available from the authors upon request.*

Contact: *jeffrey_yuan@merck.com*

In a high usage bioinformatics environment, the ability to generate high quality multiple sequence alignments with little user intervention is of great necessity. The default parameters for almost all multiple sequence alignment algorithms often results in alignments that have errors that need to be corrected either by changing the operating parameters by trial and error or by tedious hand manipulations of the alignments. Nevertheless, many types of analyses, such as phylogenetic analysis, profilesearches, and secondary structure prediction require accurate multiple sequence alignments. Therefore, to eliminate the tedium of manually determining alignment parameters and of hand tweaking alignments, we created the Perl script MULTICLUSTAL which automates the process of identifying the parameters needed to generate better multiple sequence alignments by Clustal W (Thompson *et al.*, 1994).

MULTICLUSTAL progressively changes the operating parameters of Clustal W, and scores the resulting outputs for their alignment quality and then realigns the sequences based on the best alignment parameters from the previous round of alignments. Four cycles of this alignment-scoring task are performed. The first cycle includes a reordering step, which reorders the input sequences based on the neighbor-joining tree generated from the best alignment parameters from this cycle. This step was included to alleviate the need to vary the percent identity needed to delay the multiple alignment step for certain groups of distantly related sequences. All subsequent cycles use this reordered dataset.

Six parameters are varied over the four cycles. In the first cycle, only the pairwise and multiple alignment substitution matrices are varied, the other parameters are set at default values. In the second cycle, after reordering the input dataset, the pairwise and multiple alignment gap open penalties are varied from 0 to 20 in steps of 4 using the best matrices determined in cycle one. In the third cycle, the pairwise gap extension penalties are varied from 0 to 0.5 in 0.1 intervals and the multiple gap extension penalties are varied from 0.02 to 0.1 in 0.02 intervals. Again, the best matrices and gap open penalties from the previous rounds are used. Finally, the fourth cycle reexamines the substitution matrices using the best parameters from cycles two and three. This four cycle scheme was chosen as a compromise between speed and thoroughness in examining parameter space. A total of 98 alignments are performed.

Essential to MULTICLUSTAL is the alignment quality, defined as:

$$\theta = (\iota+s)-(\gamma+(\alpha+\beta/2+\delta/4+\varepsilon/8))$$

where θ = alignment score, ι = identical amino acids, σ = conservative amino acid substitutions, γ = gap events, and where: α = # of single amino acid islands = -X-; where - is a gap and X is an amino acid; β = # of double amino acid islands = -XX-; δ = # of triple amino acid islands = -XXX-; ε = # of quadruple amino acid islands = -XXXX-.

The scoring function described above provides the highest score to those alignments that have minimized both the number of gaps and scattered 1 to 4 amino acid islands while maximizing the number of identities and similarities. Amino acid islands between one to four residues are penalized since their appearance can reflect the inappropriate introduction of gaps to maximize identities and similarities, although the penalty score for single amino acid islands is eight times higher than that for quadruple amino acid islands since longer islands can represent small conserved motifs while isolated single islands tend to reflect alignment artifacts. Islands longer than four residues are not penalized since they begin to represent the alignment of local domains.

The calculation of the alignment score requires that each alignment be piped through the Boxshade program (freely available at ftp://ftp.isrec.isb-sib.ch/pub/sib-isrec/boxshade) to generate an alignment in rich text format (RTF). This

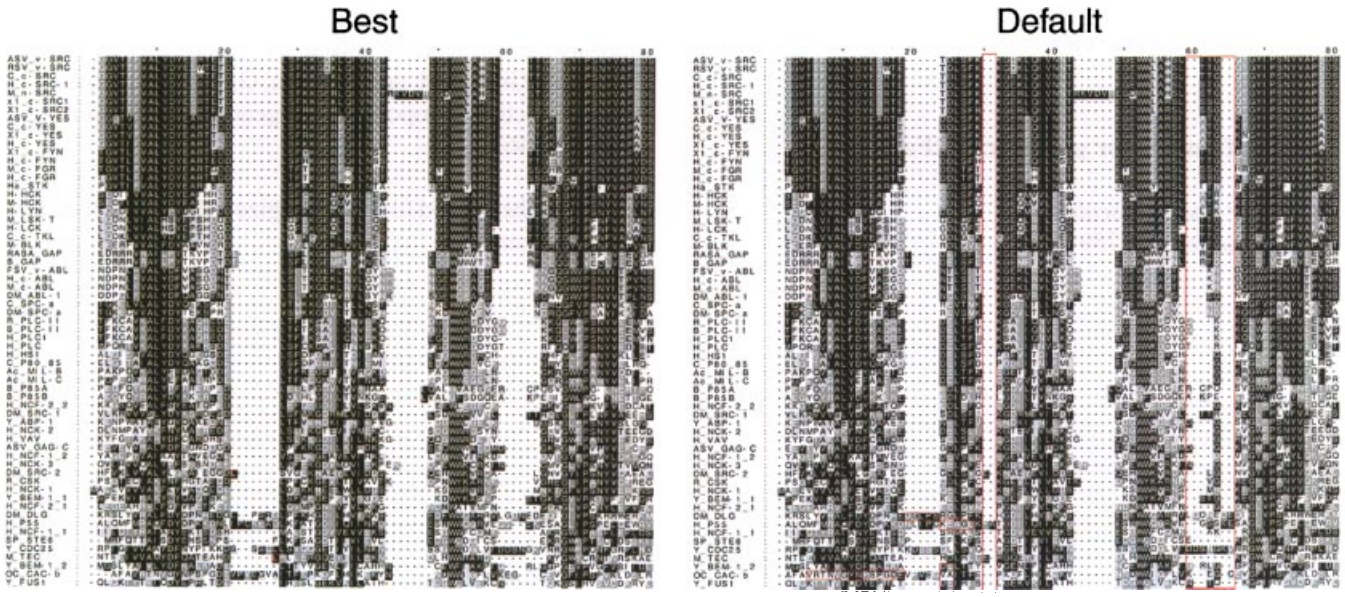

Fig. 1. Multiple sequence alignments of SH3 domain from MULTICLUSTAL and Clustal W. The SH3 domain dataset used in the original Clustal W paper (Thompson *et al.*, 1994) was first presented by Musacchio *et al.* (1992). The best alignment generated by MULTICLUSTAL is labelled 'Best'. For this dataset, MULTICLUSTAL used a PAM 350 substitution matrix with a gap open penalty of 4 and a gap extension penalty of 0 for pairwise alignments and a Blosum 30 substitution matrix with a gap open penalty of 12 and a gap extension penalty of 0.1 for multiple alignments. The alignment labeled 'Default' is generated using the default parameters of Clustal W. The alignment was shaded using the Genedoc program (freely available at http://www.cris.com/~ketchup/genedoc.shtml). The genes in the default alignment were rearranged in Genedoc after alignment to match the gene order of the MULTICLUSTAL alignment. The red boxes indicate the misaligned regions/residues. MULTICLUSTAL is implemented in Perl5 running under IRIX 6.4. A multithreaded version of Clustal W version 1.74 compiled for IRIX 6.2 and above (freely available at http://www.sgi.com/chembio/resources/clustalw/parallel_clustalw.html) was the alignment engine used by MULTICLUSTAL.

output is used to identify all the identities and similarities that occur in >20% of the aligned sequences. The RTF file along with the original ALN file from Clustal W are then scored for alignment quality by MULTICLUSTAL using the equation described above.

We tested MULTICLUSTAL on many ($n > 20$) datasets. The SH3 dataset was as a test set for MULTICLUSTAL since it was presented in the original Clustal W manuscript and its crystal structure is known, thus this dataset provided a good standard for comparison (Thompson *et al.*, 1994; Musacchio *et al.*, 1992; Yu *et al.*, 1992). The best alignment generated by MULTICLUSTAL and the default alignment from Clustal W is shown in Figure 1. Using the default parameters of Clustal W, there are at least 6 misaligned regions within the 63 SH3 domains, as shown by the outlined red boxes. The best alignment generated by MULTICLUSTAL has only two misaligned residues. The default alignment from Clustal W inserted two blocks of gaps, splitting two subdomains and creating an isolated island of glycines at the COOH terminus of the SH3 domain. The best alignment from MULTICLUSTAL did not split the two subdomains nor did it isolate the glycines into an island. Overall, the MULTICLUSTAL alignment consists of four well aligned subdomains that have fewer misaligned residues and fewer gaps than the original alignment. Finally, the MULTICLUSTAL alignment closely matches the known structural data on SH3 domains, inserting gaps only at the three known loop regions and not into the structural beta sheet subdomains (Horita *et al.*, 1998).

References

Horita,D.A., Baldisseri,D.M., Zhang,W., Altieri,A.S., Smithgall,T.E., Gmeiner,W.H. and Byrd,R.A. (1998) Solution structure of the Human Hck SH3 domain and identification of its ligand binding site. *J. Mol. Biol.*, **278,** 253–265.

Musacchio,A., Gibson,T., Lehto,V.P. and Saraste,M. (1992) SH3 – an abundant protein domain in search of a function. *FEBS*, **307,** 55–61.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22,** 4673–4680.

Yu,H., Rosen,M.K., Shin,T.B., Seidel-Dugan,C., Brugge,J.S. and Schreiber,S.L. (1992) Solution structure of the SH3 domain of Src and identification of its ligand-binding site. *Science*, **258,** 1665–1668.