BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 14 2004, pages 2309–2311
doi:10.1093/bioinformatics/bth220

MuSiC: a tool for multiple sequence alignment with constraints

Yin Te Tsai[1], Yen Pin Huang[2], Ching Ta Yu[1] and Chin Lung Lu[2,*]

[1]Department of Computer Science and Information Management, Providence University, Taiwan, ROC and [2]Department of Biological Science and Technology, National Chiao Tung University, Taiwan, ROC

Received on January 16, 2004; revised on March 5, 2004; accepted on March 5, 2004
Advance Access publication April 1, 2004

ABSTRACT

Summary: MuSiC is a web server to perform the constrained alignment of a set of sequences, such that the user-specified residues/nucleotides are aligned with each other. The input of the MuSiC system consists of a set of protein/DNA/RNA sequences and a set of user-specified constraints, each with a fragment of residue/nucleotide that (approximately) appears in all input sequences. The output of MuSiC is a constrained multiple sequence alignment in which the fragments of the input sequences whose residues/nucleotides exhibit a given degree of similarity to a constraint are aligned together. The current MuSiC system is implemented in Java language and can be accessed via a simple web interface.

Availability: http://genome.life.nctu.edu.tw/MUSIC

Contact: cllu@mail.nctu.edu.tw

INTRODUCTION

Multiple sequence alignment (MSA) has received much attention in the fields of bioinformatics and computational biology because it is very useful for discovering the biological meanings of sequences. Usually, biologists may have advanced knowledge of the structures or functionalities of sequences of interest, such as active site residues, intramolecular disulfide bonds, substrate binding sites, enzyme activities and others, as well as the conserved motifs (consensuses) of the sequences. They expect an MSA program to be able to align these sequences such that the structural/functional or conserved residues/nucleotides are aligned. However, most available MSA programs cannot do so because they generate an alignment based only on the information about the sequence. Hence, this work studies a variant of MSA, called constrained multiple sequence alignment (CMSA), which aligns the structure/function-related or conserved residues/nucleotides. Note that many other CMSAs have been proposed from various perspectives, using different approaches (Schuler *et al.*, 1991; Depiereux and Feytmans, 1992; Taylor, 1994; Notredame *et al.*, 2000; Thompson *et al.*, 2000; Katoh *et al.*, 2002).

Before the CMSA model is defined, notation must be introduced. Given an alignment $\mathcal{A}$ of k sequences $S_1, S_2, \ldots, S_k$, a band is defined as a block of consecutive columns in $\mathcal{A}$. For any band of $\mathcal{A}$, say B, $B(S_i)$ denotes the fragment of S_i whose residues/nucleotides are all in the band B, where $1 \leq i \leq k$. $d(P', P'')$ is the Hamming distance between two fragments P' and P'' of equal length. It is equal to the number of mismatched pairs in the alignment of P' and P'' without any gap. Let $l(P)$ represent the length of a fragment sequence P. The CMSA problem considered here is defined as follows. Let $\mathcal{S} = \{S_1, S_2, \ldots, S_k\}$ be a set of k input sequences along with an ordered set of r constraints $\mathcal{P} = (P_1, P_2, \ldots, P_r)$ and an error threshold ϵ, where $0 \leq \epsilon < 1$. The CMSA of $\mathcal{S}$ (with respect to $\mathcal{P}$) is an MSA of $\mathcal{S}$, say $\mathcal{A}$, in which r disjoint bands $B_1, B_2, \ldots, B_r$ are in $\mathcal{A}$ such that $d(P_i, B_i, (S_j)) \leq l(P_i) \times \epsilon$ for all $1 \leq i \leq r$ and $1 \leq j \leq k$ [meaning that at most $l(P_i) \times \epsilon$ mismatches are allowed in the alignment between P_i and $B_i(S_j)$ without any gap]. Then, the so-called CMSA problem is to find a CMSA of $\mathcal{S}$ with respect to $\mathcal{P}$ whose sum-of-pairs score is minimum over all possible candidates. Note that the original CMSA model proposed by Tang *et al.* (2003) is a special case of the model considered herein, in which each of the given constraints is a single residue/nucleotide and moreover, the aligned fragments in all bands are exactly matched to the constraint [i.e. $l(P_i) = 1$ for all $1 \leq i \leq r$ and $\epsilon = 0$]. In this simple model, a heuristic algorithm, which runs in $\mathcal{O}(rkn^4)$ time and consumes $\mathcal{O}(rn^4)$ space, was designed to find a CMSA of the input sequences, where n is the maximum of the lengths of sequences. Later, Chin *et al.* (2003) and Yu (2003) independently improved this heuristic to $\mathcal{O}(rk^2n^2)$ time and $\mathcal{O}(rn^2)$ space.

The so-called progressive method is used herein to design an algorithm of $\mathcal{O}(rk^2n^2)$ time and $\mathcal{O}(rn^2)$ space for efficiently finding a favorable CMSA of the input sequences. The method first designs a dynamic programming algorithm to find an optimal constrained alignment of any two given sequences and then use it as a kernel progressively to align

*To whom correspondence should be addressed.

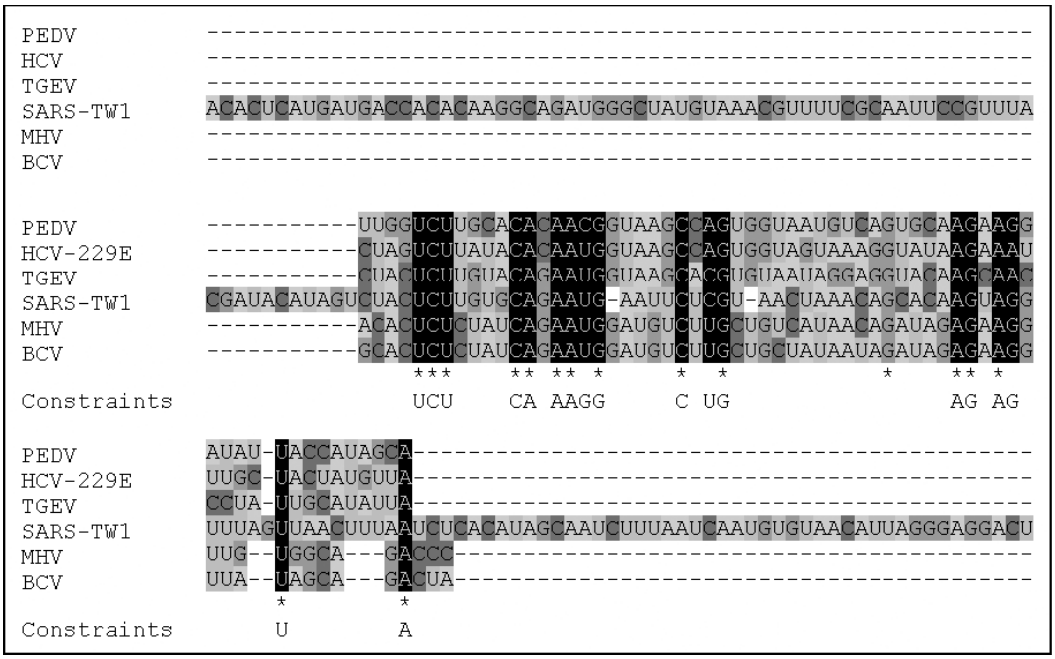

Fig. 1. The partial display of the resulting CMSA of MuSiC by aligning the sequences of SARS-TW1 3′-UTR with the pseudoknot sequences of other five coronaviruses.

the input sequences into a CMSA according to the branching order of a guide tree[1]. The algorithm was implemented as a program called MuSiC, which is a short for Multiple Sequence Alignment with Constraints. It can be easily accessed via a simple web interface. The use of the proposed MuSiC system is illustrated below to help to detect a fragment of an RNA sequence in the 3′-untranslated region (3′-UTR) of the SARS-TW1 sequence, which can fold itself into a pseudoknot structure. The structural elements in the 5′-and 3′-UTRs of a plus-strand RNA virus have been postulated to be involved in RNA replication, transcription and translation by interacting with viral or cellular proteins. Much evidence supports the fact that the pseudoknots in 3′-UTRs among coronaviruses participate in the replication of RNA (Williams *et al.*, 1999). The Severe Acute Respiratory Syndrome (SARS) virus, which caused several hundreds of deaths since its outbreak in early 2003, is a novel type of coronavirus, so a pseudoknot is expected to be observed in its 3′-UTR. By comparing the sequences of the phylogenetically conserved pseudoknots among many coronaviruses, Williams *et al.* (1999) found that these sequences contain several fragments of conserved nucleotides (consensuses). They found 11 consensuses, say `UCU`, `CA`, `AA`, `GG`, `C`, `UG`, `A`, `G`, `AG`, `U` and `A`, among coronaviruses including HCV-229E (human coronavirus), PEDV (porcine epidemic diarrhea virus), TGEV (porcine transmissible gastroenteritis virus), BCV (bovine coronavirus) and MHV (mouse hepatitis virus). To determine whether or not the 3′-UTR of SARS has a pseudoknot, SARS-TW1 (AY291451) was chosen as the test subject and the MuSiC system was used to align the sequence of the 3′-UTR of SARS-TW1 with those of the detected pseudoknots in the 3′-UTRs of BCV, MHV, PEDV, TGEV and HCV-229E coronaviruses. The consensuses described above were used as the constrained sequences in the proposed MuSiC system and then the default scoring matrix and gap penalties were chosen with the initial setting $\epsilon = 0$. As a result, no CMSA was found to satisfy the requirement, because the postulated pseudoknot in the 3′-UTR of SARS-TW1 may comprise the fragments that are only partially, rather than completely, similar to the constraints. Hence, this case was tested again with letting $\epsilon = 0.5$ so that in the band of the resulting CMSA, of length two or three, no more than one mismatch may exist between the fragment of each input sequence and the constraint. Consequently, as indicated in Figure 1, a satisfied CMSA was found. Note that the band of the resulting CMSA that corresponds to a constraint is black and its corresponding constraint is displayed beneath it. In some bands of this resulting CMSA, such as the fourth, sixth and ninth, at most one mismatch exists between the fragment of each input sequence and the corresponding constraint. Moreover, the part of SARS-TW1 aligned with the pseudoknot sequences of other coronaviruses is interspersed with only two gaps of length one. This finding suggests that this part of SARS-TW1 may fold itself into a

[1]Due to the space limitation, we leave the recursive functions and the detailed steps of our algorithm at our web site (http://genome.life.nctu.edu.tw:8080/MUSIC/method.pdf).

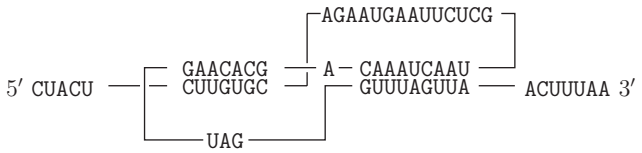

Fig. 2. The diagram of the predicted pseudoknot in the 3′-UTR of SARS-TW1 ranging from 29 460 to 29 521 bp.

pseudoknot structure and possibly be involved in replicating SARS viruses. Therefore, this SARS-TW1 fragment is further validated by applying PKNOTS, developed by the Eddy group (Rivas and Eddy, 1999), to determine whether it can fold itself into a pseudoknot structure with a stable free energy. Consequently, this fragment of SARS-TW1 indeed folds itself into a stable pseudoknot whose base pairings have a topology, as indicated in Figure 2, that is very similar to those of other coronaviruses described in the literature (Williams *et al.*, 1999). However, whether or not this SARS-TW1 fragment participates in replicating the RNA of SARS must be investigated experimentally in the laboratory.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Chuan Yi Tang for helpful suggestions while developing the algorithm and the program prototype of MuSiC, and the anonymous referees for many constructive comments in the presentation of this paper. This work was supported in part by National Science Council of Republic of China under grants NSC92-2213-E-009-089 and NSC92-3112-B-009-002.

REFERENCES

Chin,F.Y.L., Ho,N.L., Lam,T.W., Wong,P.W.H. and Chan,M.Y. (2003) Efficient constrained multiple sequence alignment with performance guarantee. *Proceedings of the Computational Systems Bioinformatics (CSB'03)*. IEEE, Los Alamitos, CA, pp. 337–346.

Depiereux,E. and Feytmans,E. (1992) MATCH-BOX: a fundamentally new algorithm for the simultaneous alignment of several protein sequences. *Comput. Appl. Biosci.*, **8**, 501–509.

Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Rivas,E. and Eddy,S. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.

Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) A workbench for multiple alignment construction and analysis. *Prot. Struct. Funct. Genet.*, **9**, 180–190.

Tang,C.Y., Lu,C.L., Chang,M.D.T., Sun,Y.J., Tsai,Y.T., Chang,J.M., Chiou,Y.H., Wu,C.M., Chang,H.T., Chou,W.I. and Chiang,S.C. (2003) Constrained sequence alignment tool development and its application to RNase family alignment. *J. Bioinform. Comput. Biol.*, **1**, 267–287.

Taylor,W.R. (1994) Motif-biased protein sequence alignment. *J. Comput. Biol.*, **1**, 297–310.

Thompson,J.D., Plewniak,F., Thierry,J.-C. and Poch,O. (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.*, **28**, 2919–2926.

Williams,G.D., Chang,R.-Y. and Brian,D.A. (1999) A phylogenetically conserved hairpin-type 39 untranslated region pseudoknot functions in coronavirus RNA replication. *J. Virol.*, **73**, 8349–8355.

Yu,C.T. (2003) Efficient algorithms for constrained sequence alignment problems. Master's Thesis, Department of Computer Science and Information Management, Providence University, Taiwan, ROC.