doi:10.1093/nar/gkm275

RE-MuSiC: a tool for multiple sequence alignment with regular expression constraints

Yun-Sheng Chung[1], Wei-Hsun Lee[2], Chuan Yi Tang[1] and Chin Lung Lu[2,3,*]

[1]Department of Computer Science, National Tsing Hua University, Hsinchu 300, Taiwan, [2]Institute of Bioinformatics, National Chiao Tung University, Hsinchu 300, Taiwan and [3]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan

Received January 31, 2007; Revised April 6, 2007; Accepted April 11, 2007

ABSTRACT

RE-MuSiC is a web-based multiple sequence alignment tool that can incorporate biological knowledge about structure, function, or conserved patterns regarding the sequences of interest. It accepts amino acid or nucleic acid sequences and a set of constraints as inputs. The constraints are pattern descriptions, instead of exact positions of fragments to be aligned together. The output is an alignment where for each pattern (constraint), an occurrence on each sequence can be found aligned together with those on the other sequences, in a manner that the overall alignment is optimized. Its predecessor, MuSiC, has been found useful by researchers since its release in 2004. However, it is noticed in applications that the pattern formulation adopted in MuSiC, namely, plain strings allowing mismatches, is not expressive and flexible enough. The constraint formulation adopted in RE-MuSiC is therefore enhanced to be regular expressions, which is convenient in expressing many biologically significant patterns like those collected in the PROSITE database, or structural consensuses that often involve variable ranges between conserved parts. Experiments demonstrate that RE-MuSiC can be used to help predict important residues and locate phylogenetically conserved structural elements. RE-MuSiC is available on-line at http://140.113.239.131/RE-MUSIC.

INTRODUCTION

Background and motivation

Sequence alignment tools are essential to biological research [see, e.g. (1), for a survey of multiple sequence alignment methods]. In addition to merely the residues/nucleotides, biologists often possess more knowledge regarding function, structure or conserved patterns of the sequences to be analyzed. It is generally desirable to have such information incorporated into an alignment procedure, so that the alignment result can be more biologically meaningful. For example, functionally important sites are generally expected to be aligned together, but a typical alignment tool often fails to achieve this if the sequence similarity is low. Imposing constraints representing such information turns out to be an effective manner to incorporate biological knowledge into an alignment tool.

Motivated by such demand, Tang *et al.* (2) formulated the constrained multiple sequence alignment problem, where each constraint is a single residue/nucleotide. They considered alignment of RNase sequences, which are known to have a sequence of conserved residues His (H), Lys (K) and His. Using H, K, H as constraints, in the resulting constrained alignment each of these three residues can be found aligned together in a column of the alignment, appearing in the order as specified. Chin *et al.* (3) then proposed an improved algorithm for pairwise alignment and an approximation algorithm for multiple alignment. It is also noted that there have been other formulations regarding alignment with constraints proposed from different perspectives with various approaches (4–14).

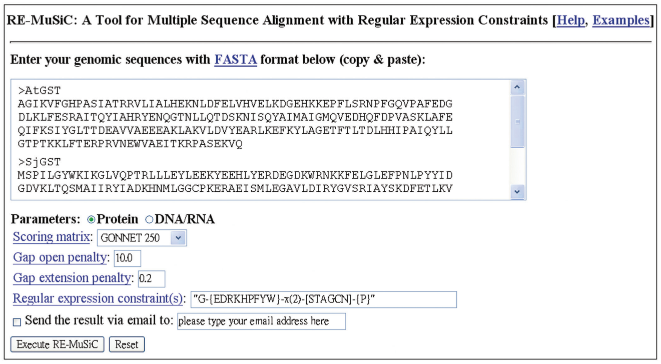
Conserved sites of a protein/RNA/DNA family are often of several residues/nucleotides long. For these patterns, the original formulation in (2) is not expressive enough. In addition, such patterns may not appear in the exact form in general. Consequently, Tsai *et al.* (15) proposed a generalized formulation and algorithm, where each constraint is a (usually short) string pattern allowing mismatches. Lu and Huang (16) then proposed a space efficient algorithm for this formulation. Web-based systems, MuSiC (15) (available at http://genome.life.nctu.edu.tw/MUSIC) and MuSiC-ME (16) (available at http://genome.life.nctu.edu.tw/MUSICME), were also developed; from now on these two systems will be referred to as MuSiC jointly. With the aid of MuSiC, Tsai *et al.* (15) and Lu and Huang (16) successfully

*To whom correspondence should be addressed. Tel: +886-3-5712121; Fax: +886-3-5729288; Email: cllu@mail.nctu.edu.tw

identified a fragment in the 3′ untranslated region (3′-UTR) of a SARS (severe acute respiratory syndrome) coronavirus sequence that can fold into a pseudoknot, which is potentially responsible for self-replication of the virus. Indeed, since its release, MuSiC has been found useful in, e.g. detection of functionally and/or structurally important residues/motifs in sequences (17,18), prediction of RNA pseudoknotted structures (15,19,20), prediction of protein structures (21) and so on.

There are, however, formulations of many biologically significant patterns beyond the capability of MuSiC. For example, many function-related protein sites as those collected in the PROSITE database (22) are expressed in regular expressions, which cannot be modeled using the substring-with-mismatch formulation of constraints implemented in MuSiC. An example of regular expression patterns is the EGF-like domain signature 2 (EGF_2, PS01186 in PROSITE): C-x-C-x(2)-[GP]-[FYW]-x(4,8)-C, which is related to the initiation of a signal transduction that results in DNA synthesis and cell proliferation. The meaning of this pattern is that, the first residue is Cys, followed by one residue of any kind, then a Cys, followed by two residues of any kind, then a Gly or Pro, etc. Regular expressions are also convenient in describing variable ranges between patterns or between blocks within a pattern, which is necessary for some single patterns themselves, and useful in applications where different patterns are expected to exhibit proximity in their occurrences. In the above example of EGF_2, the 'x(4,8)' symbol preceding the last Cys indicates a range of length varying from 4 to 8 between a residue of [F, Y or W] (Phe, Tyr or Trp) and that last Cys. Due to the usefulness of regular expressions in describing biological patterns, an enhanced web server, RE-MuSiC (*Mu*ltiple *S*equence Al*i*gnment with *R*egular *E*xpression *C*onstraints), capable of handling regular expression constraints, is developed.

DIALIGN (8,9,12,13) (http://dialign.gobics.de/) is a well-known web server that can accept user-defined constraints as anchor points. It can be noted that the constraint formulation of DIALIGN and the one of RE-MuSiC are significantly different. In DIALIGN, a constraint consists of the exact positions of a pair of equal-length segments on two of the sequences, where these two segments are expected to be aligned together. Conflicts of constraints, if any, are resolved according to a weight function defined on the segment pairs. This formulation is more similar to the one of Myers *et al.* (6). On the other hand, in RE-MuSiC, a constraint is a regular expression pattern. Each pattern may occur many times in a sequence, where each occurrence needs not have the same length. The occurrences to be aligned together so as to satisfy the constraints will be those that can make the overall alignment optimized.



RE-MuSiC: A Tool for Multiple Sequence Alignment with Regular Expression Constraints [Help, Examples]

Enter your genomic sequences with FASTA format below (copy & paste):

```
>AtGST
AGIKVFGHPASIATRRVLIALHEKNLDFELVHVELKDGEHKKEPFLSRNPFGQVPAFEDG
DLKLFESRAITQYIAHRYENQGTNLLQTDSKNISQYAIMAIGMQVEDHQFDPVASKLAFE
QIFKSIYGLTTDEAVVAEEEAKLAKVLDVYEARLKEFKYLAGETFTLTDLHHIPAIQYLL
GTPTKKLFTERPRVNEWVAEITKRPASEKVQ

>SjGST
MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYID
GDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKV
```

Parameters: ● Protein ○ DNA/RNA

Scoring matrix: GONNET 250

Gap open penalty: 10.0

Gap extension penalty: 0.2

Regular expression constraint(s): "G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}"

□ Send the result via email to: please type your email address here

Figure 1. The user-interface of RE-MuSiC. The user simply enters input sequences in FASTA format in the largest blank field. Another field with title 'Regular expression constraint(s)' is for the user to enter the regular expression constraints, in a PROSITE-like format (for details please refer to the help page at http://140.113.239.131/RE-MUSIC/help.html). In this figure, the inputs for the GST experiment are shown (see 'Experiments').

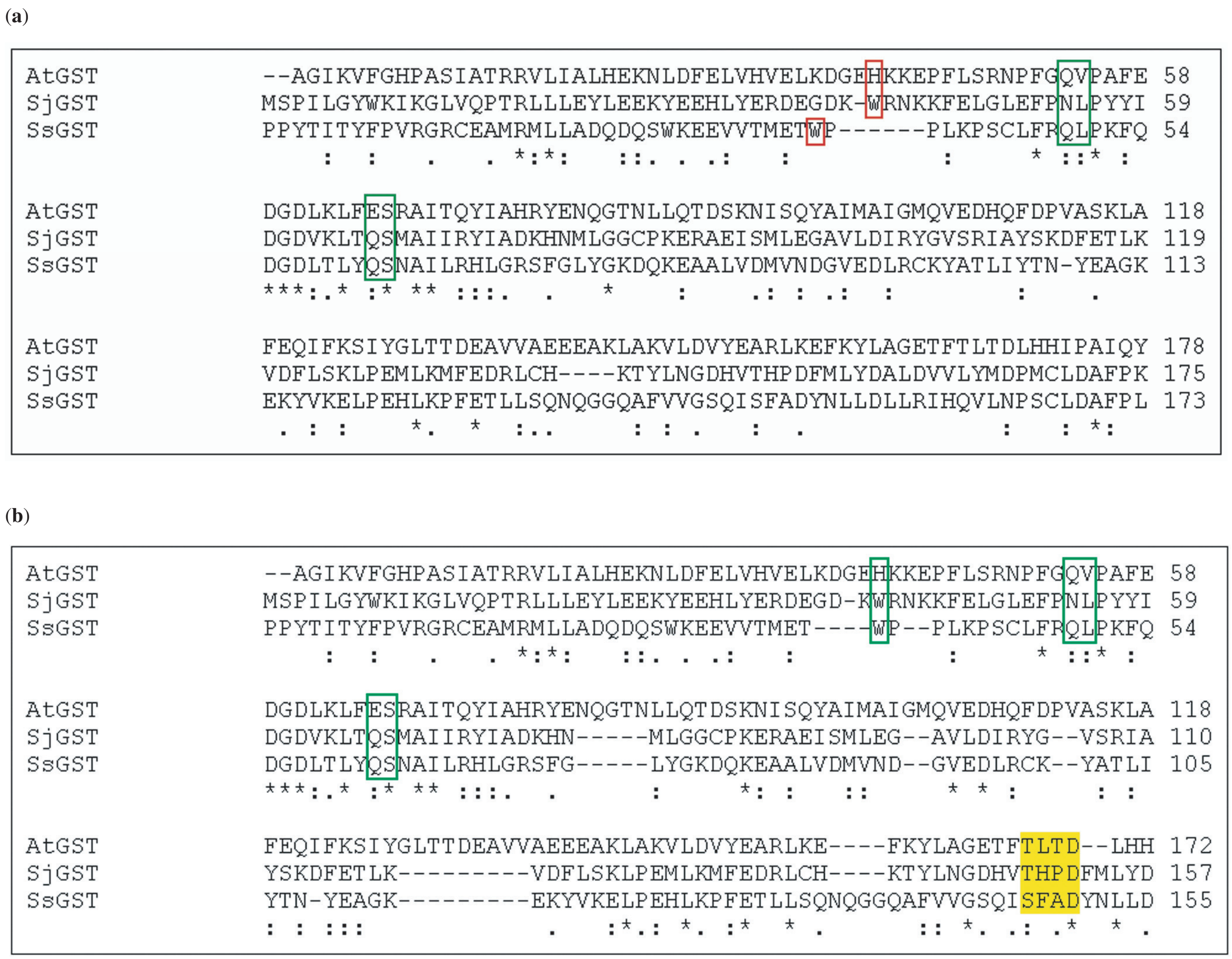
Using RE-MuSiC

RE-MuSiC provides an intuitive user interface (Figure 1). The user enters or pastes the input sequences (in FASTA format) in the largest blank field. The format for the constraints follows the PROSITE pattern format (please see the help page at http://140.113.239.131/RE-MUSIC/help.html for details). Each constraint is put within quotes, and adjacent constraints are separated by space characters. The user needs to specify whether the input sequences are proteins or DNA/RNA. The preferred scoring matrix may be chosen, and the gap open/extension penalties can be assigned. The user can also enter an email address so that a hyperlink to the alignment result will be sent via email. The output page shows the constrained alignment with the regions for the satisfactions of the constraints shaded in yellow (Figure 2b). On the output page the user can also choose to download the alignment result in FASTA format or ClustalW format.

METHODS

The regular expression constrained sequence alignment problem was originally formulated by Arslan (23). The algorithm proposed in (23) is for pairwise alignment with a single constraint. In (24) Arslan extended the algorithm in (23) to support multiple alignment with multiple constraints. The algorithm proposed in (24) may be implemented. It computes mathematically optimal constrained alignments. Unfortunately, the time complexity is extremely high, involving an exponential multiplicative factor in addition to the exponential time complexity for optimal (unconstrained) MSA computations. Even for pairwise alignment with multiple constraints, its worst case time and space requirements are intensive. In addition, the algorithms in (23,24) cannot find in the resulting alignment the regions responsible for the satisfactions of the constraints; only the alignment score, without the alignment itself, is reported. But being able to report alignments is important for a web server. It is therefore necessary to propose a solution more suitable for practical applications.

For pairwise alignment with one regular expression constraint, in a previous study (25) we have proposed an algorithm, which is more efficient both in time and in space than the one in (23). Furthermore, the alignment



(a)

```
AtGST           --AGIKVFGHPASIATRRVLIALHEKNLDFELVHVELKDGEHKKEPFLSRNPFGQVPAFE 58
SjGST           MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDK-WRNKKFELGLEFPNLPYYI 59
SsGST           PPYTITYFPVRGRCEAMRMLLADQDQSWKEEVVTMETWP------PLKPSCLFRQLPKFQ 54
                    :  :   .   . *:*:   ::. . .:   :          :     * ::* :

AtGST           DGDLKLFESRAITQYIAHRYENQGTNLLQTDSKNISQYAIMAIGMQVEDHQFDPVASKLA 118
SjGST           DGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLK 119
SsGST           DGDLTLYQSNAILRHLGRSFGLYGKDQKEAALVDMVNDGVEDLRCKYATLIYTN-YEAGK 113
                ***:.* :* ** :::.  .   *    :    .: : .:  :        :    .

AtGST           FEQIFKSIYGLTTDEAVVAEEEAKLAKVLDVYEARLKEFKYLAGETFTLTDLHHIPAIQY 178
SjGST           VDFLSKLPEMLKMFEDRLCH----KTYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPK 175
SsGST           EKYVKELPEHLKPFETLLSQNQGGQAFVVGSQISFADYNLLDLLRIHQVLNPSCLDAFPL 173
                 . : :    *.  *  :..     : : .   :  .             :   : *:
```

(b)

```
AtGST           --AGIKVFGHPASIATRRVLIALHEKNLDFELVHVELKDGEHKKEPFLSRNPFGQVPAFE 58
SjGST           MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGD-KWRNKKFELGLEFPNLPYYI 59
SsGST           PPYTITYFPVRGRCEAMRMLLADQDQSWKEEVVTMET----WP--PLKPSCLFRQLPKFQ 54
                    :  :   .   . *:*:   ::. . .:   :          :     * ::* :

AtGST           DGDLKLFESRAITQYIAHRYENQGTNLLQTDSKNISQYAIMAIGMQVEDHQFDPVASKLA 118
SjGST           DGDVKLTQSMAIIRYIADKHN-----MLGGCPKERAEISMLEG--AVLDIRYG--VSRIA 110
SsGST           DGDLTLYQSNAILRHLGRSFG-----LYGKDQKEAALVDMVND--GVEDLRCK--YATLI 105
                ***:.* :* ** :::.  .      :     *: :   ::     * * :     : :

AtGST           FEQIFKSIYGLTTDEAVVAEEEAKLAKVLDVYEARLKE----FKYLAGETFTLTD--LHH 172
SjGST           YSKDFETLK---------VDFLSKLPEMLKMFEDRLCH----KTYLNGDHVTHPDFMLYD 157
SsGST           YTN-YEAGK---------EKYVKELPEHLKPFETLLSQNQGGQAFVVGSQISFADYNLLD 155
                : : :::            .   :*.: *. :*  * .      :: *. .: .*  * .
```

Figure 2. Results of the experiment on glutathione S-transferase (GST). The boxed residues are the G-site residues shared by these GSTs. (**a**) A partial view of the alignment produced by ClustalW. Not all of the G-site residues can be aligned together. (**b**) A partial view of the alignment produced by RE-MuSiC. A common pattern of these GST sequences, namely '[ST]-x(2)-[DE]' (PS00006), as obtained from PROSITE, is used as constraint. In the resulting alignment, RE-MuSiC annotated the region for the satisfaction of the constraint with a yellow block. It can be seen that the G-site residues are aligned properly, as desired. For both tools, the default parameter settings are adopted.

in addition to the score can be reconstructed without worsening the time and space complexity. In this work we extend the algorithm in (25) to support multiple constraints and multiple sequences, as required in RE-MuSiC. The resulting algorithm is more efficient than the one in (24) for pairwise alignment with multiple constraints. To deal with multiple sequences, a progressive method is implemented, using our improved pairwise algorithm as the kernel. For details of the algorithm the reader is referred to the supplementary material (available at http://140.113.239.131/RE-MUSIC/RE_MuSiC_method.pdf).

EXPERIMENTS

Protein sequences with active site residues

The glutathione binding site (G-site) on glutathione S-transferase (GST) had been found to have conserved architectures across species (26). The chemical natures of their residues acting as G-site ligands and interactions facilitated with glutathione are also analogous (26). In a reasonable alignment of GST protein sequences, therefore, the residues for the G-site are expected to be aligned together. A structural superposition of the crystal structures of GST proteins from different species also suggests that most of these G-site residues should be aligned together (26). The sequence identity of those GST proteins from different species, however, is quite low; for example, it is reported in (26) that the pairwise sequence identity between the *A. thaliana* GST and each of other six non-plant GSTs is no more than 20.2%. In such a case, interfered by the low-similarity regions, it would be difficult for a typical alignment tool to align the important residues well. An experiment is therefore undertaken to examine the performance of a typical alignment tool in this case, as well as to demonstrate how RE-MuSiC can be used to produce a more reasonable alignment.

In this experiment we analyze three GST proteins: (i) AtGST: a phi class GST from plant *A. thaliana*

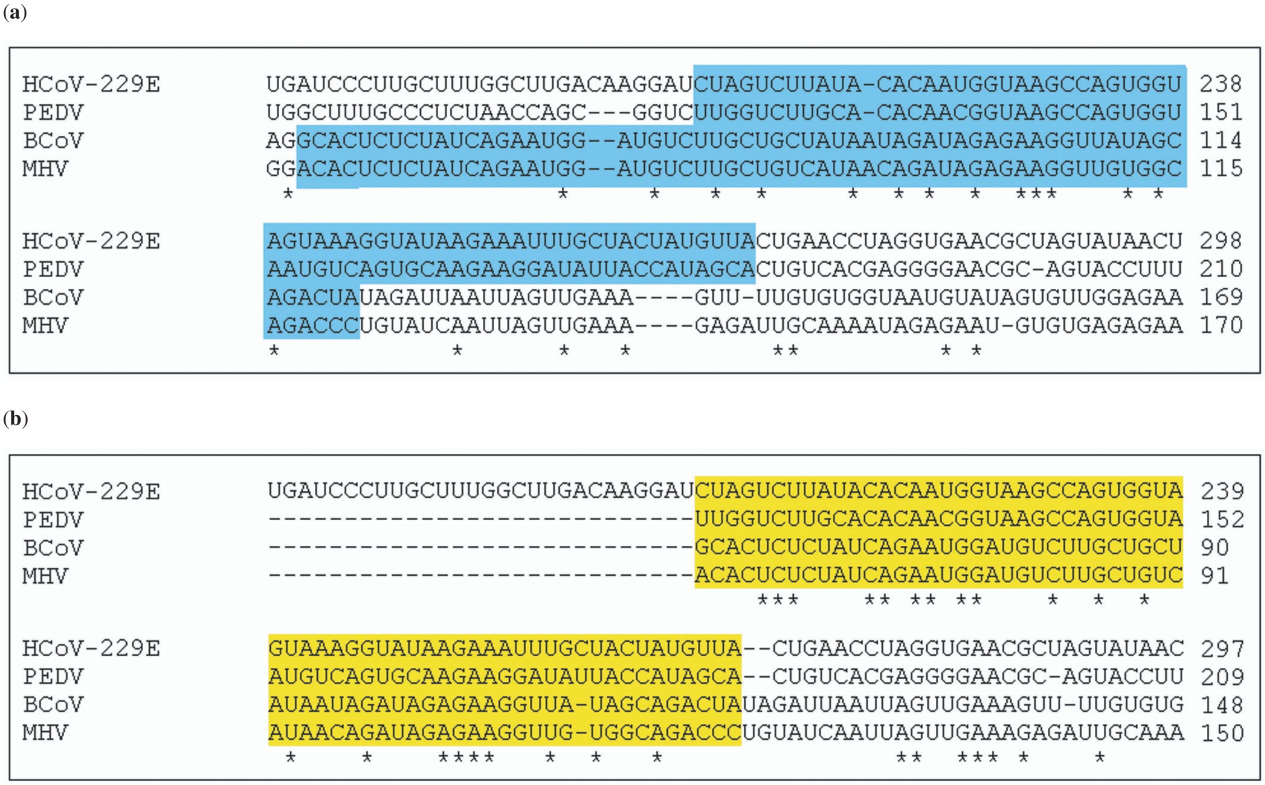
(PDBID: 1GNW); (ii) SjGST: an alpha class GST from non-mammalian *S. japonicum* (flat worm) (PDBID: 1M99); (iii) SsGST: a pi class GST from mammalian *S. scrofa* (pig) (PDBID: 2GSR). These sequences are first aligned using ClustalW (27). The result is shown in Figure 2a, where active site residues shared by these GSTs are boxed. It can be seen that, part of the G-site residues failed to be aligned together, due to the low sequence similarity among these GST proteins. By querying PROSITE with the three proteins, it is found that they all share the pattern PS00006 ([ST]-x(2)-[DE]). Using this pattern as constraint, RE-MuSiC is applied to align the sequences again. The result is shown in Figure 2b. As expected, the common pattern is aligned together. Meanwhile, the G-site residues are aligned properly, as desired. These suggest that, with some information about common patterns, RE-MuSiC is more reliable to produce alignments in which biologically important residues can be lined up, which is particularly important when the sequence identity is low. Being more reliable in aligning together important residues, RE-MuSiC may also be applied to align an unknown sequence with other sequences whose relevant residues are known, thus providing a convenient and cheap way for a preliminary prediction of the residues in question on the unknown sequence. Note also that, in this experiment, the knowledge about the active site residues are not utilized in constructing the alignment; the constraints do not involve the active site residues themselves. Such a property is useful when the residues to be predicted are not expected to be conserved in the sequence level.

RNA sequences with phylogenetically conserved pseudoknots

There is considerable evidence that suggests phylogenetically conserved pseudoknots found in the 3′-UTRs of various coronaviruses are involved in RNA replication of these viruses (28). In an alignment of the 3′-UTR sequences of coronaviruses, therefore, it is desirable if these pseudoknots can be aligned together. However, it is often the case that the sequence identity among the coronaviruses from different groups is low. It is not an easy task for a typical alignment tool to align together the conserved pseudoknots. In this experiment, we demonstrate that RE-MuSiC can be helpful in this situation.

Four coronaviruses are considered in this experiment (GenBank accession numbers in parentheses): (i) HCoV-229E: human 229E coronavirus (af304460), (ii) PEDV: porcine epidemic diarrhea virus (af353511), (iii) BCoV: bovine coronavirus (af220295) and (iv) MHV: murine hepatitis virus (af201929). The first two are group 1 coronaviruses, while the others belong to group 2.

First, ClustalW is applied to align these coronavirus sequences. The result is shown in Figure 3a. Not surprisingly, since the sequence identity is low,



(a)

```
HCoV-229E       UGAUCCCUUGCUUUGGCUUGACAAGGAUCUAGUCUUAUA-CACAAUGGUAAGCCAGUGGU 238
PEDV            UGGCUUUGCCCUCUAACCAGC---GGUCUUGGUCUUGCA-CACAACGGUAAGCCAGUGGU 151
BCoV            AGGCACUCUCUAUCAGAAUGG--AUGUCUUGCUGCUAUAAUAGAUAGAGAAGGUUAUAGC 114
MHV             GGACACUCUCUAUCAGAAUGG--AUGUCUUGCUGUCAUAACAGAUAGAGAAGGUUGUGGC 115
                 *                 *     *   *  *     *  * *  *  ***    * *

HCoV-229E       AGUAAAGGUAUAAGAAAUUUGCUACUAUGUUACUGAACCUAGGUGAACGCUAGUAUAACU 298
PEDV            AAUGUCAGUGCAAGAAGGAUAUUACCAUAGCACUGUCACGAGGGGAACGC-AGUACCUUU 210
BCoV            AGACUAUAGAUUAAUUAGUUGAAA----GUU-UUGUGUGGUAAUGUAUAGUGUUGGAGAA 169
MHV             AGACCCUGUAUCAAUUAGUUGAAA----GAGAUUGCAAAAUAGAGAAU-GUGUGAGAGAA 170
                *           *      *   *         **         * *
```

(b)

```
HCoV-229E       UGAUCCCUUGCUUUGGCUUGACAAGGAUCUAGUCUUAUACACAAUGGUAAGCCAGUGGUA 239
PEDV            ----------------------------UUGGUCUUGCACACAACGGUAAGCCAGUGGUA 152
BCoV            ----------------------------GCACUCUCUAUCAGAAUGGAUGUCUUGCUGCU 90
MHV             ----------------------------ACACUCUCUAUCAGAAUGGAUGUCUUGCUGUC 91
                                                ***    ** ** **    *  *  *

HCoV-229E       GUAAAGGUAUAAGAAAUUUGCUACUAUGUUA--CUGAACCUAGGUGAACGCUAGUAUAAC 297
PEDV            AUGUCAGUGCAAGAAGGAUAUUACCAUAGCA--CUGUCACGAGGGGAACGC-AGUACCUU 209
BCoV            AUAAUAGAUAGAGAAGGUUA-UAGCAGACUAUAGAUUAAUUAGUUGAAAGUU-UUGUGUG 148
MHV             AUAACAGAUAGAGAAGGUUG-UGGCAGACCCUGUAUCAAUUAGUUGAAAGAGAUUGCAAA 150
                 *    *    ****   *  *   *               **  *** *    *
```

Figure 3. Results of the experiment on coronaviruses with 3′-UTR pseudoknots. (**a**) A partial view of the alignment produced by ClustalW. The shaded regions, corresponding to the phylogenetically conserved pseudoknots, are not aligned well. (**b**) A partial view of the alignment produced by RE-MuSiC. The consensus of the pseudoknots on the four coronaviruses involves variable ranges between residues. RE-MuSiC has a constraint formulation flexible enough to express this consensus. As expected, the regions for the pseudoknots are aligned properly by RE-MuSiC. For both tools, the default parameter settings are adopted.

the phylogenetically conserved pseudoknots (shaded regions) are not aligned well. In (28), predicted secondary structures of the pseudoknots found in the 3′-UTR of various coronaviruses are given. A consensus of the pseudoknots is to be taken. Since, in general, loops in pseudoknots are less conserved, to enhance flexibility, we exclude loop regions nucleotides from the consensus. Then the consensus of the pseudoknots can be described as 'x(5)-C-U-x(4)-C-x(15,16)-U-G-x(2)-A-x(5,7)-G-x(4)-A-G-x(7,10)-U-x(3)-A-x(5).' Using this consensus as the constraint, RE-MuSiC is applied to align these 3′-UTR sequences again. In Figure 3b, the pseudoknot regions on these coronaviruses can be seen to have been aligned properly. This demonstrates that RE-MuSiC can be used to help locate fragments that are conserved in structure. Actually, this property, being a common advantage of the MuSiC series, had been utilized to predict the pseudoknot in the 3′-UTR of the SARS-TW1 coronavirus by aligning the 3′-UTR of SARS-TW1 with those of some other coronaviruses whose pseudoknot regions are known (15,16). RE-MuSiC further makes it possible to provide the flexibility of variable ranges between conserved nucleotides or regions in constraints, which is necessary for describing the whole consensus of the pseudoknot in this experiment. This is a significant advance over previous generations of MuSiC.

SUMMARY

Imposing constraints is an effective manner to incorporate biological knowledge into an alignment tool. Previous versions of MuSiC do not support many biologically significant patterns. RE-MuSiC adopts regular expressions as its constraint formulation, which is useful in expressing PROSITE patterns or structural elements that often involve variable ranges between conserved parts. The algorithm underlying RE-MuSiC represents an improvement over the previously proposed algorithm, and is more appropriate for implementation in a web-server. Experiments on GST proteins and on coronaviruses with phylogenetically conserved pseudoknots demonstrate that, with additional knowledge incorporated, RE-MuSiC is able to produce meaningful alignments in which important residues or structural elements can be aligned properly, even if the similarity among input sequences is low. Such ability is also useful for prediction purposes.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for their useful comments, which improved the usability of the web interface and the clarity of this article. This work was supported in part by National Science Council of Republic of China under grant NSC95-2221-E-009-231 (C.L. Lu) and grants NSC95-2221-E-007-031 and NSC95-2627-B-007-002 (C.Y. Tang), and in part by the ATU plan of MOE. Funding to pay the Open Access publication charges for this article was provided by the ATU plan of MOE.

Conflict of interest statement. None declared.

REFERENCES

1. Notredame,C. (2002) Recent progresses in multiple sequence alignment: a survey. *Pharmacogenomics*, **3**, 131–144.
2. Tang,C.Y., Lu,C.L., Chang,M.D.T., Tsai,Y.T., Sun,Y.J., Chao,K.M., Chang,J.M., Chiou,Y.H., Wu,C.M. *et al.* (2003) Constrained multiple sequence alignment tool development and its application to RNase family alignment. *J. Bioinform. Comput. Biol.*, **1**, 267–287.
3. Chin,F.Y.L., Ho,N.L., Lam,T.W., Wong,P.W.H. and Chan,M.Y. (2005) Efficient constrained multiple sequence alignment with performance guarantee. *J. Bioinform. Comput. Biol.*, **3**, 1–18.
4. Shuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) A workbench for multiple alignment construction and analysis. *Proteins: Struct. Funct. Genet.*, **9**, 180–190.
5. 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.
6. Myers,G., Selznick,S., Zhang,Z. and Miller,W. (1996) Progressive multiple alignment with constraints. *J. Comput. Biol.*, **3**, 563–572.
7. Morgenstern,B., Dress,A. and Werner,T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA*, **93**, 12098–12103.
8. Morgenstern,B., Frech,K., Dress,A. and Werner,T. (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290–294.
9. Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
10. Thompson,J.D., Plewniak,F., Thierry,J.-C. and Poch.,O. (2000) DbClustal: rapid and reliable multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.*, **28**, 2919–2926.
11. Sammeth,M., Morgenstern,B. and Stoye,J. (2003) Divide-and-conquer multiple alignment with segment-based constraints. *Bioinformatics*, **19**(Suppl. 2), ii189–ii195.
12. Morgenstern,B. (2004) DIALIGN: Multiple DNA and protein sequence alignment at BiBi-Serv. *Nucleic Acids Res.*, **32**, W33–W36.
13. Morgenstern,B., Werner,N., Prohaska,S.J., Schneider,R.S.I., Subramanian,A.R., Stadler,P.F. and Weyer-Menkhoff,J. (2005) Multiple sequence alignment with user-defined constraints at GOBICS. *Bioinformatics*, **21**, 1271–1273.
14. Morgenstern,B., Prohaska,S.J., Pohler,D. and Stadler,P.F. (2006) Multiple sequence alignment with user-defined anchor points. *Algorithms Mol. Biol.*, **1**, 6.
15. Tsai,Y.-T., Huang,Y.P., Yu,C.T. and Lu,C.L. (2004) MuSiC: a tool for multiple sequence alignment with constraints. *Bioinformatics*, **20**, 2309–2311.
16. Lu,C.L. and Huang,Y.P. (2005) A memory-efficient algorithm for multiple sequence alignment with constraints. *Bioinformatics*, **21**, 20–30.
17. Song,B., Choi,J.H., Chen,G.Y., Szymanski,J., Zhang,G.Q., Tung,A.K.H., Kang,J., Kim,S. and Yang,J. (2006) ARCS: an aggregated related column scoring scheme for aligned sequences. *Bioinformatics*, **22**, 2326–2332.
18. Cheng,C.Y., Chang,C.H., Wu,Y.J. and Li,Y.K. (2006) Exploration of glycosyl hydrolase family 75, a chitosanase from Aspergillus fumigatus. *J. Biol. Chem.*, **281**, 3137–3144.
19. Huang,C.H., Lu,C.L. and Chiu,H.T. (2005) A heuristic approach for detecting RNA H-type pseudoknots. *Bioinformatics*, **21**, 3501–3508.
20. Reeder,J., Hochsmann,M., Rehmsmeier,M., Voss,B. and Giegerich,R. (2006) Beyond Mfold: Recent advances in RNA bioinformatics. *J. Biotechnol.*, **124**, 41–55.
21. Dunbrack,R.L. (2006) Sequence comparison and protein structure prediction. *Curr. Opin. Struct. Biol.*, **16**, 374–384.
22. Hulo,N., Sigrist,C.J.A., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, 134–137.

23. Arslan,A.N. (2005) Regular expression constrained sequence alignment. In *Proceedings of 16th Annual Symposium on Combinatorial Pattern Matching (CPM05)*, Vol. 3537 of *Lecture Notes in Computer Science*, Springer, Berlin, pp. 322–333.
24. Arslan,A.N. (2005) Multiple sequence alignment containing a sequence of regular expressions. In *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB05)*. San Diego, USA, pp. 1–7.
25. Chung,Y.-S., Lu,C.L. and Tang,C.Y. (2006) Efficient algorithms for regular expression constrained sequence alignment. In *Proceedings of 17th Annual Symposium on Combinatorial Pattern Matching (CPM06)*, Vol. 4009 of *Lecture Notes in Computer Science*, Springer, Berlin, pp. 389–400.
26. Reinemer,P., Prade,L., Hof,P., Neuefeind,T., Huber,R., Zettl,R., Palme,K., Schell,J., Koelln,I. *et al.* (1996) Three-dimensional structure of glutathione S-transferase from *Arabidopsis thaliana* at 2.2 Å resolution: structural characterization of herbicide-conjugating plant glutathione S-transferases and a novel active site architecture. *J. Mol. Biol.*, **255**, 289–309.
27. 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.
28. Williams,G.D., Chang,R.Y. and Brian,D.A. (1999) A phylogenetically conserved hairpin-type 3′ untranslated region pseudoknot functions in coronavirus RNA replication. *J. Virol.*, **73**, 8349–8355.