BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 12 2007, pages 1553–1555
doi:10.1093/bioinformatics/btm128

Sequence analysis

SQUINT: a multiple alignment program and editor

Matthew G. Goode and Allen G. Rodrigo*

Bioinformatics Institute, Private Bag 92019, University of Auckland, Auckland, New Zealand

Received on December 6, 2006; revised on February 8, 2007; accepted on March 26, 2007

Advance Access publication May 7, 2007

Associate Editor: Keith Crandall

ABSTRACT

Summary: SQUINT is a sequence alignment tool, and combines both automated progressive sequence alignment with facilities for manual editing. The program imports nucleotide or amino acid sequence multiple alignment files in standard formats, and permits users to view two translations of the same multiple alignment simultaneously. Edits in one view are instantaneously reflected in the other, and the scoring cost of the changes are shown in real-time. Progressive multiple alignments, using a variety of alignment parameters, can be performed on any block of sequences, including blocks embedded in the existing alignment.

Availability: The software is freely available for download at http://www.cebl.auckland.ac.nz

Contact: a.rodrigo@auckland.ac.nz

A multiple sequence alignment is a matrix of nucleotide or amino acid sequences that have been aligned such that each column in the matrix contains only those sites that share a common evolutionary history (i.e. the sites in any single column are homologous). If, over the course of evolution, a block of sites in a given sequence has been deleted or inserted, a multiple alignment needs to reflect the insertion/deletion (indel) event by including gaps that indicate the absence of homologous sites in the sequence in question (in the case of a deletion) or in other sequences (in the case of an insertion). Multiple sequence alignment can be performed automatically using a variety of existing programs and online resources (Edgar, 2004; Poirot *et al*., 2003; Thompson *et al*., 1994). Usually, however, these computer-generated alignments (CGAs) have to be edited manually. There are five main reasons for this:

(1) Most CGAs operate using a heuristic approach and, therefore, are not guaranteed to find an optimal solution; consequently, manual editing can frequently improve the alignment;

(2) CGAs may not preserve biologically important motifs or domains, and these may need to be re-assembled;

(3) CGAs performed on one sequence type may provide unsatisfactory alignments when translated into an alternative sequence type (e.g. CGAs of protein-coding nucleotide sequences may introduce frameshifts and/or stop-codons);

(4) Some algorithms used for constructing CGAs use the same alignment parameters along the entire length of the alignment whereas, in reality, different regions of the alignment may require more or less stringency with respect to the creation and extension of indels.

(5) Finally, for downstream analyses, it may be necessary to edit CGAs to remove blocks of sequences that may be ambiguously aligned, or may have too many indels.

SQUINT (SeQuence alIgnmeNT) is a multiple alignment program that has both heuristic multiple sequence alignment and alignment editing facilities. The program can import unaligned or aligned nucleotide or amino acid sequences in FASTA (Limpman and Pearson, 1985), CLUSTAL (Thompson *et al*., 1994), PHYLIP (Felsenstein, 1996), and NEXUS (Maddison *et al*., 1997) formats, and can export nucleotide or amino acid alignments in these same formats. Standard alignment editing functions are available in SQUINT. These include gap-insertion and deletion, and the exclusion of some sites or sequences. If protein-coding nucleotide sequences are imported, sequences can be reversed (with or without base complementation), the first codon may be specified, and sequences can be translated into amino acid sequences using any one of 14 genetic codes and viewed/manipulated as such. Sequences can also be viewed as codons, and the standard edits can also be performed on the codon alignment.

Protein-coding nucleotide sequence alignments can also be viewed as 'gap-balanced' alignments (Rodrigo *et al*., 2000), in which the placement of the first, second and third positions of each codon along a sequence is displayed as '[X]', where 'X' is the encoded amino acid, and the left and right parentheses denote the first and third codon positions, respectively. A gap-balanced alignment permits a user to detect nucleotide sites that do not appear to align to the correct codon position. We have found this facility useful in detecting potential sequencing errors, particularly erroneous insertions or deletions by the base-calling algorithm. Similarly, if the aim is to generate alignments for evolutionary or phylogenetic analysis of selection using ratios of non-synonymous to synonymous substitutions, it is imperative that coding nucleotide sequences are properly aligned with respect to their codon positions.

If sequences are unaligned, SQUINT can perform a standard progressive alignment (Feng and Doolittle, 1987), although SQUINT is not meant as a replacement for a full-featured alignment program. The guide tree used in SQUINT is a UPGMA tree constructed using the scores of all pairwise alignments. Standard alignment scoring matrices,

*To whom correspondence should be addressed.

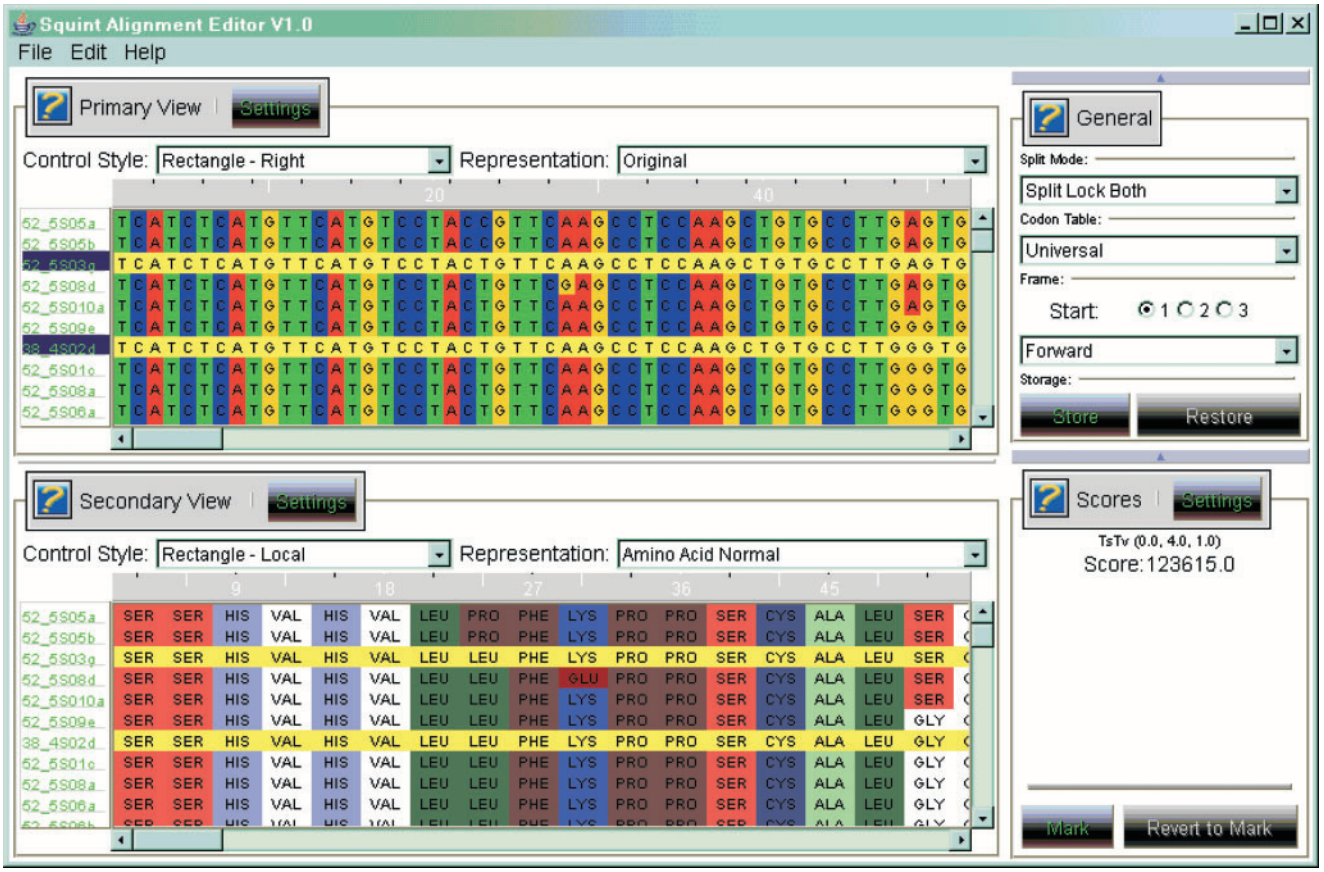

Fig. 1. The main window of the Squint application, showing a split view of a currently loaded alignment. One view shows the nucleotide alignment, while an amino acid alignment is shown in the other. There is one currently active scoring metric.

including PAM250 (Dayhoff *et al*., 1978) and BLOSUM62 (Henikoff and Henikoff, 1992), are available, and users are also able to define their own matrices for nucleotide sequences. A novel feature of SQUINT is the ability to align codon sequences. As with other standard progressive alignment programs, users can also specify the cost of creating a gap in an alignment, the cost of extending a gap, and the cost of gaps at the start or end of an alignment (end-gaps).

A key feature of the progressive alignment facility is the ability to perform a partial alignment, whereby selected segments of an existing alignment are aligned independently. Segments can be a block of sequences, a block of sequential columns, or even more refined blocks of selected sequences and columns. Any scoring matrix may be used to do this, thus permitting the user to realign different regions of an existing alignment using different parameters. Once realigned, the selected segment is re-embedded in the existing overall alignment. This is done by treating both the realigned segment and the unaltered altered section as two alignments, then realigning both in the manner of a standard progressive alignment.

The score(s) of an alignment using one or more available scoring matrices can also be viewed by the user. This feature is particularly useful, because the scoring function in SQUINT is updated in real-time as the alignment is edited. This allows the user to determine whether an edit or a partial alignment has improved the multiple alignment. The score can be viewed as an absolute score, and as a score relative to a previously 'marked' alignment.

A significant feature of SQUINT is the ability to view alignments in two windows. Each window can have a different translation of the alignment; for instance, if protein-coding nucleotide sequences are available, a user can view the original nucleotide alignment as well as the translated amino-acid alignment (Fig. 1). Edits and partial alignments can be performed in either window, and changes in one window are simultaneously shown in the other. This means that users can perform a progressive amino acid alignment on protein-coding nucleotide sequences, and retrieve both the amino acid and nucleotide multiple alignments. The two-window facility permits users to determine how edits on one datatype (i.e. nucleotides or amino acids) affects the translated alignment.

SQUINT is written in Java 1.5, and is available for computers with the Windows, Linux and MacOS X operating systems.

ACKNOWLEDGEMENTS

M.G.G. was supported by a scholarship from the Allan Wilson Centre for Molecular Ecology and Evolution.

The authors would also like to thank Howard Ross and Lea Deng, for their useful comments in the development of SQUINT. A.G.R. completed this work while on sabbatical at Olivier Gascuel's laboratory in the Laboratoire d'Informatique, de Microélectronique et de Robotique de Montpellier.

Conflict of Interest: none declared.

REFERENCES

Dayhoff,M.O. *et al*. (1978) A model for evolutionary change in proteins. In Dayhoff,M.O. (ed.) *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Silver Spring, MD, Vol. 5, Suppl. 3, pp. 345–352.

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

Felsenstein,J. (1996) Phylip. Copyright: University of Washington. http://evolution.genetics.washington.edu/phylip.html

Feng,D.F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol*., **25**, 351–360.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Lipman,D.J. and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.

Maddison,D.R. *et al*. (1997) Nexus: an extensible file format for systematic information. *Syst. Biol*., **46**, 590–621.

Poirot,O. *et al*. (2003) Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res*., **31**, 3503–3506.

Rodrigo,A.G. *et al*. (2000) Sampling and processing HIV molecular sequences: a computational evolutionary biologist's perspective. In Rodrigo, A. G. and Learn,G.H. (eds) *The Computational Evolutionary Analysis of HIV Molecular Sequences*. Wolters-Kluwer Academic Publishers, Boston.

Thompson,J.D. *et al*. (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.