Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

DNATagger, colors for codons

N.M. Scherer[1] and D.M. Basso[2]

[1]Bioinformatics Department, Heinrich-Heine-Universität, Düsseldorf, Germany
[2]Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

Corresponding author: N.M. Scherer
E-mail: scherer@cs.uni-duesseldorf.de

Genet. Mol. Res. 7 (3): 853-860 (2008)
Received June 2, 2008
Accepted August 11, 2008
Published September 16, 2008

ABSTRACT. DNATagger is a web-based tool for coloring and editing DNA, RNA and protein sequences and alignments. It is dedicated to the visualization of protein coding sequences and also protein sequence alignments to facilitate the comprehension of evolutionary processes in sequence analysis. The distinctive feature of DNATagger is the use of codons as informative units for coloring DNA and RNA sequences. The codons are colored according to their corresponding amino acids. It is the first program that colors codons in DNA sequences without being affected by "out-of-frame" gaps of alignments. It can handle single gaps and gaps inside the triplets. The program also provides the possibility to edit the alignments and change color patterns and translation tables. DNATagger is a JavaScript application, following the W3C guidelines, designed to work on standards-compliant web browsers. It therefore requires no installation and is platform independent. The web-based DNATagger is available as free and open source software at http://www.inf.ufrgs.br/~dmbasso/dnatagger/.

Key words: Protein-coding DNA; Codon alignment; Bioinformatics tool; Alignment colorization; Visualization and editing

INTRODUCTION

Genes are made of DNA (or RNA in some organisms) and carry the information for constructing whole organisms. The genetic sequences evolve over time, accumulating mutations that account for the differentiation and diversity of living beings. The comparative analysis of homologous genes provides helpful insights into the evolutionary history of genes and organisms.

Protein-coding genes are translated into amino acid polypeptides following the genetic code. The DNA sequence of a gene directly determines the sequence of amino acids in the protein it produces (Lewin, 1996). Following a determined reading frame, each group of three consecutive nucleotides in the DNA (or RNA) sequence corresponds to an amino acid residue that will be incorporated into the protein sequence. These nucleotide triplets are called "codons". The correspondence between the codons and their coded amino acids constitutes the genetic code.

With an alphabet of four letters representing nucleotides, there are 64 triplet arrangements. From these 64 codons, up to four, depending on each genetic code, are nonsense (stop codons), and they signal the end of the translation. The standard genetic code has three stop codons, while there are examples of alternative codes having one (e.g., alternative flatworm mitochondrial code), two (e.g., yeast mitochondrial code) or four stop codons (e.g., vertebrate mitochondrial code). Each of the 60 (or more) sense codons specifies one of the 20 amino acids. Consequently, each amino acid can be coded by more than one triplet. This redundancy is referred to as the "degeneracy" of the genetic code. This property allows the occurrence of silent mutations, i.e., mutations that alter the DNA sequence, while the coded protein remains with the same amino acid sequence. The information about silent mutations can be very informative in phylogenetic and evolutionary studies (Goldman and Yang, 1994). Protein-coding nucleotide sequences carry evolutionary information that is hidden in their amino acid counterparts. However, the four bases present in the DNA will not reflect the diversity of the residues in the protein (Bininda-Emonds, 2005). Thus, it is practical to combine information from amino acid and DNA sequences.

Two or more related sequences can be compared in an alignment. They are arranged in a matrix and compared in order to find similarities and differences. The sequences being compared are expected to be homologous, i.e., derived from a common ancestor. Each single sequence is placed as a row, whereas columns represent homologous sites. Over the evolutionary history, DNA sequences accumulate mutations, which can be nucleotide substitutions, insertions or deletions. To compensate insertions and deletions ("indels") in an alignment, the introduction of "gaps" is necessary (Needleman and Wunsch, 1970), normally represented by dashes. It is important to point out that, although they represent evolutionary events, gaps do not belong to the genomic sequence, and therefore, they do not interfere in the translation process.

Coloring the characters representing nucleotide or amino acid residues in multiple sequence alignments makes their visualization more agreeable and understandable. Various coloring schemes have been developed to highlight residue conservation, amino acid properties, nucleotide types, and even structural patterns (Parry-Smith et al., 1998; Beitz, 2000). Commonly used software for alignment editing, visualization and presentation, such as ClustalX (Thompson et al., 1997), CINEMA (Parry-Smith et al., 1998) and TEXshade (Beitz, 2000), shade each character or each column in an alignment independently.

Only a few programs for alignment editing and visualization consider codons and their correspondence to amino acids in order to color DNA sequences. MacClade (Maddison and Maddison, 2000) and Se-Al (Rambaut, 1996-2002) are only available for Macintosh machines. However, these programs also assume a model of evolution where nucleotide insertions and deletions always occur in multiples of three and always start and stop at codon boundaries. The requirement that indels always have to start and stop at codon boundaries does not reflect biological reality.

Guided by the need for visualization of the coding character of DNA sequences, we developed a tool for coloring nucleotide alignments on the basis of their translated sequences. DNATagger is the first program that colors codons in DNA sequences without being affected by "out-of-frame" gaps in alignments. It can handle single gaps and gaps inside the triplets. The coloring of codons allows frame-shift mutations to be easily recognized in the alignment. DNATagger is dedicated to the visualization of protein coding sequences and also protein sequence alignments to facilitate the comprehension of evolutionary processes in sequence analysis. It also provides the possibility to edit the alignments and change color patterns and translation tables (Figure 1). The simple text editing function provides freedom over the standard formats required by most applications.

The DNATagger web interface can be found at http://www.inf.ufrgs.br/~dmbasso/dnatagger/ and it is free and open source software (FOSS).

MATERIAL AND METHODS

Software description

DNATagger is designed to colorize codons in DNA/RNA sequence alignments, and amino acids in protein sequence alignments (Figure 1). Therefore, all DNA/RNA sequences and sequence alignments are interpreted as being from protein coding genes. The coding frame is determined by the first base letter recognized in each sequence, irrespective of its position in the alignment. Gaps occurring in the sequence string are interpreted as non-informative characters, and are not included as part of the codon. For example, the sequence string `ACAT--GG-C-T-A` will be displayed with the colors representing the amino acids coded by the codons `ACA`, `TGG`, and `CTA`. The gaps will continue to be displaced in the original position, but they will not be colorized.

Sequences and alignments should be introduced as plain text on the "Edit" tab by pasting or typing. There is no need for file uploads. The user can work with his/her alignments in raw text. The simple text editing function provides freedom over the standard formats required by most applications. The sequences are interpreted with respect to the genetic code chosen by the user. There are 17 genetic codes available at the selection box on the "Translation Table" tab. By default, the standard code is used.

Amino acid and codon sequences are colored on the "View" tab following the color table selected by the user on the "Colors" tab. Several color combinations are already available. Figure 1 shows the Taylor table, constructed according to Taylor (1997). A monochromatic table was also created, motivated by the publication costs of color illustrations. The user is invited to create his/her own color table using the color bar option, or to write his/her own code. This table can be saved for future use in a text file. For reuse of this color table, the user should place this code in the code box and click on "interpret this color table code".

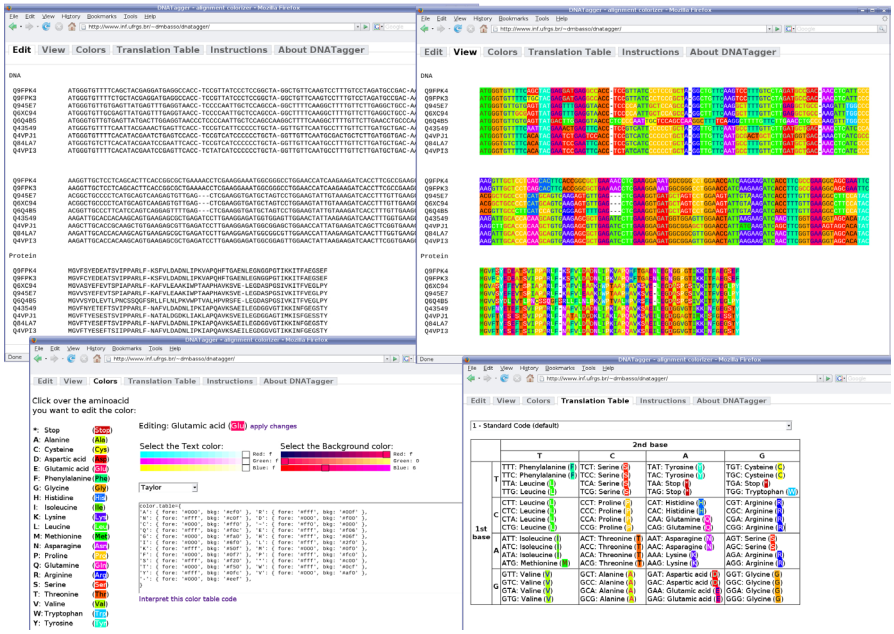

Figure 1. Screenshots from DNATagger running on Firefox. Top left, the "Edit" tab with examples of DNA and protein alignments. Top right, the "View" tab with colored sequences. Bottom left, the "Color" tab, and bottom right, the "Translation Table" tab.

Implementation details

The algorithm begins searching for codon blocks and amino acid blocks anywhere in the source text, and the colorization proceeds in accordance with the color and translation tables selected by the user. The identification of codon and amino acid blocks is made by regular expressions. An amino acid block is defined as a word boundary, followed by an amino acid letter, and then at least 3 letters from the amino acid alphabet (Σ = {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}) or gaps ('-'), followed by a word boundary. A codon block is slightly more complex. It is defined as a word boundary followed by a letter from the nucleotide alphabet (Σ = {A,C,G,T,U}), and then a nucleotide letter or gap ('-' or '.'), at least seven of any nucleotide letters, gaps or spaces, followed by a word boundary. One disadvantage is that in a few cases some words or identifiers may be wrongly understood as an alignment section. We suggest the use of identifiers containing at least one number or character that are not present in the amino acid alphabet to avoid such situations. The advantage of this method is the independence of formats, since the sequences can be arbitrarily placed in the text. It enables the colorization of most of the common alignment formats, as well as simple placed sequences.

For the colorization of codons, each triplet of nucleotide letters in the DNA or RNA sequence, beginning with the first letter of the sequence, is compared to the translation table and colored according to the amino acid it represents. Gaps or spaces inside the sequences are counted out and will not interfere in the interpretation of the coding properties. In the amino acid sequences, each amino acid letter is colored independently.

The JavaScript infrastructure used in the web-based version of DNATagger improves the user's experience, providing extended functions such as color table selection and editing. It also enables both on-line and off-line execution, the former with the application being retrieved from the Internet, and the latter having the web browser load it from a local file. To create this local file, it is only necessary to save the complete web page from the on-line address. There is no need for any installation process.

The interface is coded as a regular HTML page associated with a CSS (Cascading Style Sheet, a W3C standard technology), with several elements connected to the program's core using event handlers. The program is composed of four objects, each encapsulating the methods to deal with menu selection, color table editing and selection, translation table selection, translation, and alignment colorization. The communication between the interface and the application's objects is shown in Figure 2. Almost all interactions back to the user are made through the manipulation of the DOM (Document Object Model) tree.

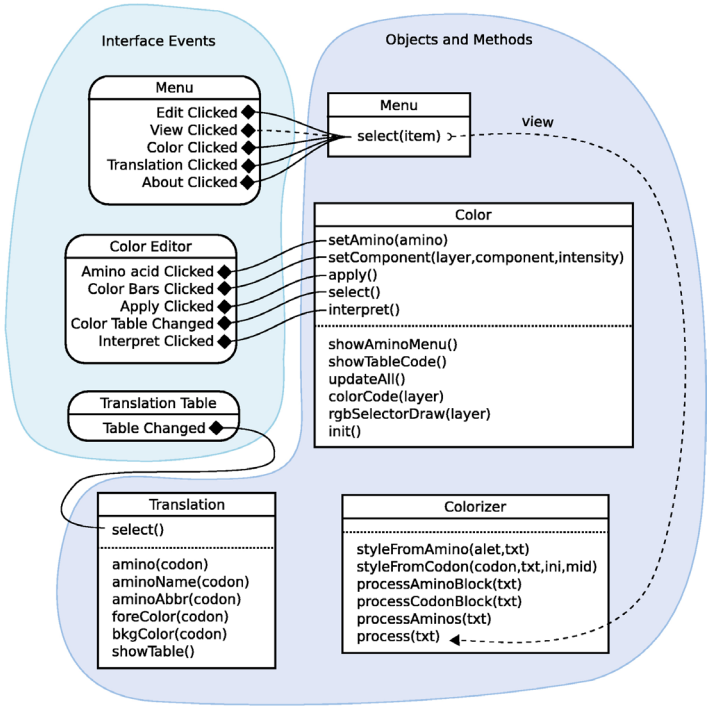

Figure 2. Communication between the interface and the application's objects.

The first version of DNATagger was written in C++ and designed as a command line application. It accepts strictly Clustal/aln format for the input file, with restriction on the lengths of the sequence names. The output is an HTML file with the colored sequences also in Clustal/aln format. The advantage of this method is the possibility to integrate it as part of a script. The disadvantage of the first version is the format restriction. The sequence names must have exactly six characters and must be followed by exactly 10 blank spaces. We are now improving

this script to allow other formats, as in the web version. It will soon be available on request.

RESULTS AND DISCUSSION

Comparison with other tools

In recent years, diverse programs, web servers and web-based applications have been improved to provide researchers in molecular biology and related areas with powerful tools for the analysis of molecular data. DNATagger is a helpful application to visualize and improve the results of coding DNA/RNA and peptide alignments. The differential feature of this program is its interpretation of nucleotide sequences as protein coding, i.e., triplets of bases as codons. Particularly important is its capability to work with "out-of-frame" gaps, interpreting each sequence in the alignment independently. Other programs that offer the possibility of coloring codons with respect to the coded amino acid cannot handle gaps that occur inside codon boundaries. The proprietary software MacClade (Maddison and Maddison, 2000) and the freely distributed software Se-Al (Rambaut, 1996-2002) have this property. Both offer a data editor tool with an option for protein-coding data that uses the position of the nucleotides in the alignment to determine the reading frame. Gaps inside the codons are treated as ambiguous characters that are part of the codon itself. This causes an artificial shift in the reading frame, and the rest of the sequence is misinterpreted until a complementary number of gaps compensate the error (Figure 3).

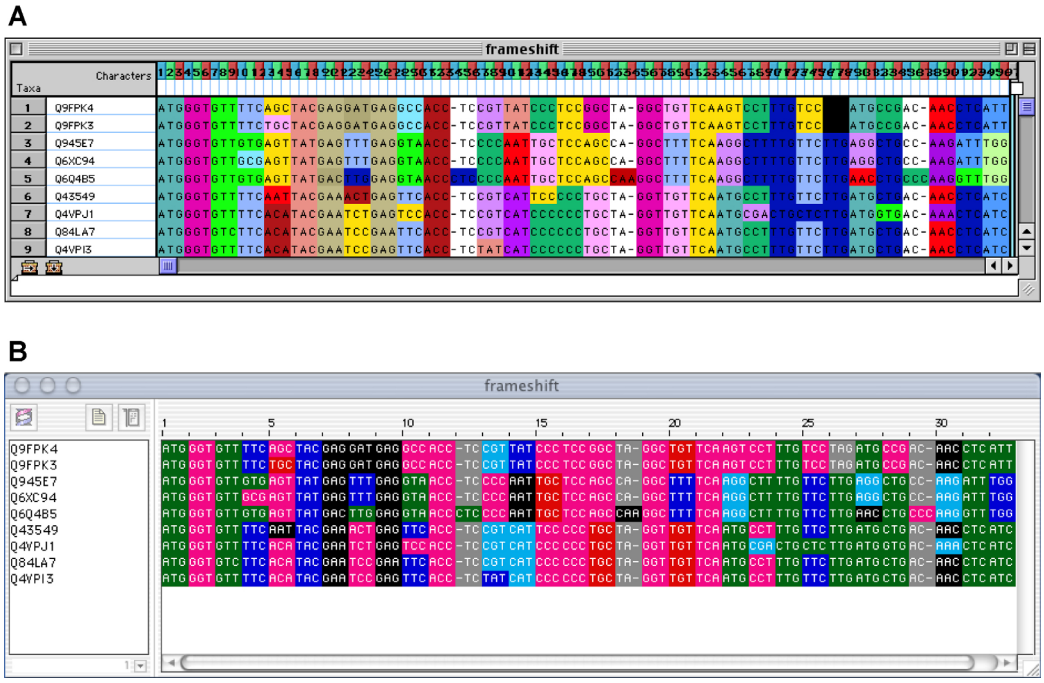

Figure 3. Partial alignment of PR-10 sequences. **A.** Displayed by MacClade version 4.0 PPC (Maddison and Maddison, 2000). **B.** Displayed by Se-Al v2.0a11 Carbon (Rambaut, 1996-2002).

Manual comparisons between nucleotide- and amino acid-based alignments help in the identification of frame-shift mutations in the sequences. Some frame-shift mutations are very clear. The example of the partial alignment of PR-10 (Pathogenesis-related Protein 10, data from Scherer NM, unpublished results) sequences in Figure 4 shows a balanced frame-shift mutation that can only be recognized in the DNA alignment. The most often used approach to align nucleotide sequences of protein coding genes is the "back-translation" of amino acid alignments (Wernersson and Pedersen, 2003). Comparing the nucleotide alignment (Figure 4A) with the alignment of the translated sequences (Figure 4B), it is clear why the amino acid alignment will not reflect the real evolutionary history of these sequences by the use of a "back-translation" approach. The reason is that the codon positions as triplets in the amino acid-based alignment are no longer really homologous. In a study of molecular evolution based on codon models (for example, Goldman and Yang, 1994), the inclusion of these "artificially aligned" codons will not represent the actual history of the site, and is therefore not compatible with the analysis. In this case, it is recommended to exclude that sequence or this excerpt of the alignment from the analysis.

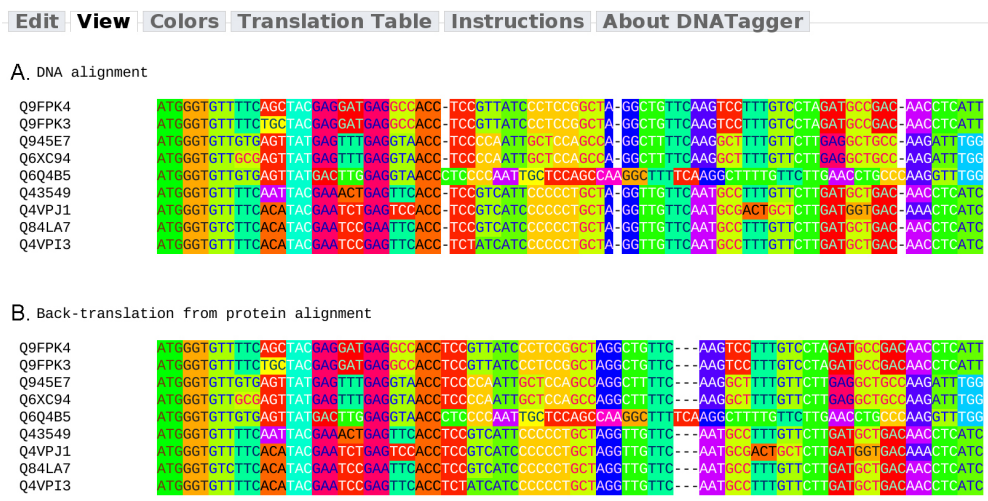

Figure 4. Partial alignment of PR-10 sequences displayed by DNATagger. **A.** DNA alignment from ClustalX (Thompson et al., 1997) shows a balanced frame-shift mutation. **B.** RevTrans (Wernersson and Pedersen, 2003) back-translation of an amino acid alignment from ClustalX. The nucleotides have been shifted to fit the amino acid positions. Their positions do not correspond to the homologous nucleotides anymore.

CONCLUSIONS

The use of DNATagger does not require installation and it is platform independent. It is available as an item of web-based free and open source software (FOSS). The entire process occurs in the browser of the user's computer and, once loaded, works off-line. It does not need to communicate with the server, and there is no file download or upload. The next step is to implement a function to export the colored alignments to a file for use in presentations and publications.

Availability and requirements

The DNATagger web interface can be found at http://www.inf.ufrgs.br/~dmbasso/dnatagger/. We highly recommend the use of a "standards-compliant" web browser, such as Mozilla Firefox, Google Chrome, Konqueror or Opera. The application is platform independent, having been developed under Linux and tested under Mac OS and Windows. A command line version of the program can be requested at dnatagger@gmail.com.

ACKNOWLEDGMENTS

We thank Ana L.C. Bazzan for supporting the idea, users and colleagues for tests and helpful suggestions, Nelson J.R. Fagundes and Rosvita Schreiner for many discussions, Claudio Scherer and Robert Warren for corrections to the manuscript, and the Institute of Informatics at UFRGS for hosting DNATagger. We also thank two anonymous reviewers for useful comments on an earlier version of the manuscript. N.M. Scherer was the recipient of a scholarship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

Beitz E (2000). TEXshade: shading and labeling of multiple sequence alignments using LATEX2 epsilon. *Bioinformatics* 16: 135-139.

Bininda-Emonds OR (2005). transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics* 6: 156.

Goldman N and Yang Z (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11: 725-736.

Lewin R (1996). Patterns in Evolution - The New Molecular View. Scientific American Library, New York.

Maddison DR and Maddison WP (2000). MacClade Version 4: Analysis of Phylogeny and Character Evolution. Sinauer Associates, Sunderland.

Needleman SB and Wunsch CD (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48: 443-453.

Parry-Smith DJ, Payne AW, Michie AD and Attwood TK (1998). CINEMA - a novel colour INteractive editor for multiple alignments. *Gene* 221: GC57-GC63.

Rambaut A (1996-2002). Sequence Alignment Editor. Version 2.0a11 Carbon. University of Oxford, Oxford. Available at http://evolve.zoo.ox.ac.uk/.

Taylor WR (1997). Residual colours: a proposal for aminochromography. *Protein Eng.* 10: 743-746.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, et al. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25: 4876-4882.

Wernersson R and Pedersen AG (2003). RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.* 31: 3537-3539.

