*Nucleic Acids Research, 2006, Vol. 34, Web Server issue* ***W609–W612***
*doi:10.1093/nar/gkl315*

# PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments

**Mikita Suyama[1], David Torrents[1] and Peer Bork[1,2,*]**

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany and
[2]Max Delbrück Center for Molecular Medicine, D-13092 Berlin-Buch, Germany

Received April 3, 2006; Revised April 6, 2006; Accepted April 11, 2006

## ABSTRACT

**PAL2NAL is a web server that constructs a multiple codon alignment from the corresponding aligned protein sequences. Such codon alignments can be used to evaluate the type and rate of nucleotide substitutions in coding DNA for a wide range of evolutionary analyses, such as the identification of levels of selective constraint acting on genes, or to perform DNA-based phylogenetic studies. The server takes a protein sequence alignment and the corresponding DNA sequences as input. In contrast to other existing applications, this server is able to construct codon alignments even if the input DNA sequence has mismatches with the input protein sequence, or contains untranslated regions and polyA tails. The server can also deal with frame shifts and inframe stop codons in the input models, and is thus suitable for the analysis of pseudogenes. Another distinct feature is that the user can specify a subregion of the input alignment in order to specifically analyze functional domains or exons of interest. The PAL2NAL server is available at http://www.bork.embl.de/pal2nal.**

## INTRODUCTION

An increasing body of research is based on the classification and evaluation of the rate of DNA evolution in coding regions. In many cases nucleotide substitutions are classified according to their impact on the encoded protein and the resulting classification can be used for a variety of analyses. Classification of synonymous ($K_S$) and non-synonymous ($K_A$) substitutions can be used to detect the presence or absence of selection (1) and the classification of substitutions according to codon position can be used with sophisticated evolutionary models to better reconstruct phylogenies. In both of these cases the programs that perform the analysis, such as PAML (2), require a codon alignment as input.

Owing to the degeneracy of the genetic code, where a single amino acid can be encoded by multiple codons, it is often preferable to align protein sequences rather than the underlying coding DNA as it increases sensitivity at longer evolutionary distances and prevents the introduction of frame shifts into an alignment. Thus the construction of a protein alignment first and then reverse translating this into a codon-based DNA alignment is invariably the optimal solution and provides reliable alignments to perform correct evolutionary analyses. In the ideal case where the protein and the corresponding DNA match perfectly, the conversion from a protein alignment into the corresponding codon alignment can be achieved by replacing each amino acid residue with three nucleotide residues, a procedure which is implemented in several tools, e.g. aa_to_dna_aln in the Bioperl toolkit (3), RevTrans (4), transAlign (5) and aa2dna (http://www.bio.psu.edu/People/Faculty/Nei/Lab/aa2dna.zip). However it often happens that the corresponding DNA sequence has mismatches when compared with its theoretical protein sequence, or contains untranslated regions (UTRs) and polyA tails. In such instances, the conversion process is much more complicated. Moreover the analysis of pseudogenes, which are an interesting subject of molecular evolution studies, requires dealing with frame shifts and inframe stop codons. These situations, which are rather frequent in large-scale analysis of sequenced genomes, cannot be solved by the programs mentioned above and therefore require additional solutions.

Here we describe a web server, PAL2NAL, which converts a protein sequence alignment into the corresponding codon alignment, despite the presence of mismatches between the protein and the DNA sequences, UTRs and polyA tails in the input DNA sequences, and frame shifts and inframe stop codons in the input alignment. Another useful feature of this server is that it is possible to obtain codon alignments for specific regions of interest, such as functional domains or

*To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 8517; Email addresses: bork@embl.de

A

```
CLUSTAL W multiple sequence alignment

BC070280        MVGSLNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQNLVIMGKKTWFSIPEKNRPLKGRINLVLSR
pseudogene      ----LNCIVNVSQKMGIIRNGDLP*PQLKNKF2-FQRMTTPSSAEGKENLVFLIRKNWFSITEKNQPLKYIINLVVSR
                    ##################             ##########################################

BC070280        ELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMLWIVGGSSVYKEAMNHPGHLKLFVTRIMQDFESDTFF-PEIDLE
pseudogene      ESKEPPQRPPFLD*SLGDALKRIEQLKLANKQDVFFTVGGSSVYKESMN*-DHFKLFVTWIMQDFQSDTFFS4EGDLE
                             ######################

BC070280        KYKLLPEYP-GVLSDVQEEKGIKYKFEVYEKND
pseudogene      KYKLLPEYPQGVVSDVEEEKGIKYKFEVYEKND
```

B

```
>BC0700280  dihydrofolate reductase (human)
TGTAACGAGC GGGCTCGGAG GTCCTCCCGC TGCTGTCATG GTTGGTTCGC TAAACTGCAT CGTCGCTGTG TCCCAGAACA TGGGCATCGG
CAAGAACGGG GACCTGCCCT GGCCACCGCT CAGGAATGAA TTCAGATATT TCCAGAGAAT GACCACAACC TCTTCAGTAG AAGGTAAACA
GAATCTGGTG ATTATGGGTA AGAAGACCTG GTTCTCCATT CCTGAGAAGA ATCGACCTTT AAAGGGTAGA ATTAATTTAG TTCTCAGCAG
AGAACTCAAG GAACCTCCAC AAGGAGCTCA TTTTCTTTCC AGAAGTCTAG ATGATGCCTT AAAACTTACT GAACAACCAG AATTAGCAAA
TAAAGTAGAC ATGcTCTGGA TAGTTGGTGG CAGTTCTGTT TATAAGGAAG CCATGAATCA CCCAGGCCAT CTTAAACTAT TTGTGACAAG
GATCATGCAA GACTTTGAAA GTGACACGTT TTTTCCAGAA ATTGATTTGG AGAAATATAA ACTTCTGCCA GAATACCCAG GTGTTCTCTC
TGATGTCCAG GAGGAGAAAG GCATTAAGTA CAAATTTGAA GTATATGAGA AGAATGATTA ATATGAAGGT GTTTTCTAGT TTAAGTTGTT
CCCCCTCCCT CTGAAAAAAG TATGTATTTT TACATTAGAA AAGGTTTTTT GTTGACTTTA GATCTATAAT TATTTCTAAG CAACTTGTTT
TTATTCCCCA CTACTCTTGT CTCTATCAGA TACCATTTAT GAGACATTCT TGCTATAACT AAGTGCTTCT CCAAGACCCC AACTGAGTCC
CCAGCACCTG CTACAGTGAG CTGCCATTCC ACACCCATCA CATGTGGCAC TCTTGCCAGT CCTTGACATT GTCGGGCTTT TCACATGTTG
GTAATATTTA TTAAAGATGA AGATCCACAT ACCCTTCAAA AAAAAAAAAA AAAAAAAA AAAAAAA
>pseudogene  dihydrofolate reductase pseudogene (human)
CTAAACTGCA TTGTCAATGA TTCCCAGAAG ATGGGCATCA TCAGGAATGG GGACCTGCCC TGACCTCAGC TCAAAAATAA ATTCGATTCC
AAAGAATGAC CACACCCTCT TCAGCAGAGG GTAAAGAAAA TTTAGTATTT TTAATTAGGA AGAACTGGTT CTCGATTACT GAGAAGAATC
AACCTTTAAA GTATATAATT AATTTAGTTG TCAGTAGAGA ATCCAAGGAA CCACCGCAAA GACCTCCTTT TCTTGACTAA AGTCTGGGTG
ATGCCTTAAA ACGTATTGAG CAACTAAAAT TAGCAAATAA ACAAGACGTG TTTTTTACAG TGGGAGGCAG TTCTGTTTAT AAGGAATCCA
TGAATTGAGA CCATTTTAAA CTATTTGTGA CATGGATCAT GCAGGACTTT CAAAGTGACA CGTTTTTTTC CCCTAGAAGG TGATTTAGAG
AAATATAAAC TTCTCCCAGA ATACCCACAA GGTGTTGTCT CTGATGTGGA GGAGGAGAAA GGCATTAAGT ACAAATTTGA AGTATATGAA
AAGAATGAT
```

C

```
#-------------------------------------------------------------------------#
#  WARNING: pseudogene pepAlnPos 11: V does not correspond to GAT
#-------------------------------------------------------------------------#

CLUSTAL W multiple sequence alignment

BC070280        ATGGTTGGTTCGCTAAACTGCATCGTCGCTGTGTCCCAGAACATGGGCATCGGCAAGAACGGGGACCTGCCCTGGCCA
pseudogene      ------------CTAAACTGCATTGTCAATGATTCCCAGAAGATGGGCATCATCAGGAATGGGGACCTGCCCTGACCT
                            ##################################################

BC070280        CCGCTCAGGAATGAATTCAGATATTTCCAGAGAATGACCACAACCTCTTCAGTAGAAGGTAAACAGAATCTGGTGATT
pseudogene      CAGCTCAAAAATAAATTCGA----TTCCAAAGAATGACCACACCCTCTTCAGCAGAGGGTAAAGAAAATTTAGTATTT
                                      ##################################################

BC070280        ATGGGTAAGAAGACCTGGTTCTCCATTCCTGAGAAGAATCGACCTTTAAAGGGTAGAATTAATTTAGTTCTCAGCAGA
pseudogene      TTAATTAGGAAGAACTGGTTCTCGATTACTGAGAAGAATCAACCTTTAAAGTATATAATTAATTTAGTTGTCAGTAGA
                ##################################################################

BC070280        GAACTCAAGGAACCTCCACAAGGAGCTCATTTTCTTTCCAGAAGTCTAGATGATGCCTTAAAACTTACTGAACAACCA
pseudogene      GAATCCAAGGAACCACCGCAAAGACCTCCTTTTCTTGACTAAAGTCTGGGTGATGCCTTAAAACGTATTGAGCAACTA
                                                                  ############################

BC070280        GAATTAGCAAATAAAGTAGACATGCTCTGGATAGTTGGTGGCAGTTCTGTTTATAAGGAAGCCATGAATCACCCAGGC
pseudogene      AAATTAGCAAATAAACAAGACGTGTTTTTTACAGTGGGAGGCAGTTCTGTTTATAAGGAATCCATGAATTGA---GAC
                ####################################

BC070280        CATCTTAAACTATTTGTGACAAGGATCATGCAAGACTTTGAAAGTGACACGTTTTTT---CCA---GAAATTGATTTG
pseudogene      CATTTTAAACTATTTGTGACATGGATCATGCAGGACTTTCAAAGTGACACGTTTTTTTCCCCTA--GAAGGTGATTTA


BC070280        GAGAAATATAAACTTCTGCCAGAATACCCA---GGTGTTCTCTCTGATGTCCAGGAGGAGAAAGGCATTAAGTACAAA
pseudogene      GAGAAATATAAACTTCTCCCAGAATACCCACAAGGTGTTGTCTCTGATGTGGAGGAGGAGAAAGGCATTAAGTACAAA


BC070280        TTTGAAGTATATGAGAAGAATGAT
pseudogene      TTTGAAGTATATGAAAAGAATGAT
```

D

```
   2    234
BC070280
CTAAACTGCATCGTCGCTTCCCAGAACATGGGCATCGGCAAGAACGGGTTCCAGAGAATGACCACAACCTCTTCA
GTAGAAGGTAAACAGAATCTGGTGATTATGGGTAAGAAGACCTGGTTCTCCATTCCTGAGAAGAATCGACCTTTA
AAGGGTAGAATTAATTTAGTTGATGCCTTAAAACTTACTGAACAACCAGAATTAGCAAATAAAGTAGACATGCTC
TGGATAGTT
pseudogene
CTAAACTGCATTGTCAATTCCCAGAAGATGGGCATCATCAGGAATGGGTTCCAAAGAATGACCACACCCTCTTCA
GCAGAGGGTAAAGAAAATTTAGTATTTTTAATTAGGAAGAACTGGTTCTCGATTACTGAGAAGAATCAACCTTTA
AAGTATATAATTAATTTAGTTGATGCCTTAAAACGTATTGAGCAACTAAAATTAGCAAATAAACAAGACGTGTTT
TTTACAGTG
```

**Figure 1.** An example of PAL2NAL input and output files. (**A**) The first input file: A multiple sequence alignment of human dihydrofolate reductase (GenBank accession no. BC070280) and its pseudogene in the CLUSTAL format with the notation used in GeneWise for frame shifts. Frame shifts and inframe stop codons in the pseudogene are shown in orange. Under the alignment, arbitrarily selected blocks are specified with '#'. (**B**) The second input file: The corresponding DNA (or mRNA) sequences in the FASTA format. UTRs and polyA tails are shown in cyan to indicate how these regions are excluded from the resulting output. (**C**) Output with the default option setting. The position of the codon that does not correspond with the input protein sequence is shown in red. The regions of alignment blocks correspond to those specified in the input protein alignment are indicated by '#'. (**D**) Output with the following option setting: *Remove mismatches*, yes; *Use only selected positions*, yes; *Output format*, PAML. With this setting, the codon alignment corresponding to the specified regions is generated in the PAML format.

particular exons by selecting the positions in the input protein sequence alignment.

## METHODS AND IMPLEMENTATION

The server requires a multiple sequence alignment of proteins and the corresponding DNA sequences as input. The internal action of the program can be divided into three main steps: (i) upload the protein sequence alignment and DNA sequences, (ii) reverse translation, i.e. conversion of the protein sequences into the corresponding DNA sequences in the form of regular expression patterns and (iii) generation of the codon alignment. In the second step, each protein sequence is converted into DNA sequence of a regular expression. For example, a short peptide sequence, MDP, is reverse-translated into a regular expression pattern of the DNA sequence as `(A(U|T)G)(GA(U|T|C|Y))(CC.)`. For frame shifts, we adapted the notation used in GeneWise (6): if an insertion or deletion is found in the coding region, it is represented by the number of nucleic acid residues at that site instead of an amino acid code. For example, M2P indicates that there is 1 nt deletion between methionine and proline. With this notation, it is easy to convert the peptide sequence into a regular expression pattern, in this case `(A(U|T)G)..(CC.)`. After converting into a regular expression pattern, the input DNA sequence is searched with the pattern to obtain the corresponding coding region. Unmatched DNA sequence regions are discarded. The pattern matching has been designed to be tolerant of mismatches. This was achieved by extending 10 amino acid regular expression matches in both directions until the entire coding region of the input DNA sequence is covered. The regions between the extended fragments and those not covered by the extension are taken as mismatches, and reported, if any, in the output. In the third step, the protein sequence alignment is converted into the corresponding codon alignment by replacing each amino acid residue with the corresponding codon sequence.

## USAGE

The PAL2NAL server takes the following two files as input: (i) a multiple sequence alignment of proteins either in the CLUSTAL or in the FASTA format and (ii) the corresponding DNA sequences in the FASTA format. An example of the application of PAL2NAL is shown in Figure 1. In this example, a multiple sequence alignment of human dihydrofolate reductase in the CLUSTAL format (7) (Figure 1A) and the corresponding DNA (or mRNA) sequences in the FASTA format (Figure 1B) are used as input. The second sequence of this example is a pseudogene, and it contains two frame shifts and three inframe stop codons. The program automatically trims all UTRs and polyA tails, and successfully converts the protein alignment into the codon alignment, despite the presence of a mismatch and two frame shifts (Figure 1C). If some positions of the input alignment are marked with '#' under the alignment (Figure 1A), the corresponding regions are also marked in the output codon alignment (Figure 1C).

There are six options in the PAL2NAL server. (i) *Codon tables*. The users can select a codon table: either 'universal' (default) or 'vertebrate mitochondrial'. (ii) *Remove gaps and inframe stop codons*. This option excludes codons with gaps and inframe stop codons from the output. This option should be selected if the codon alignment is to be further analyzed by codeml to calculate $K_S$ and $K_A$ since the PAML package does not accept codon alignments containing gaps or inframe stop codons. (iii) *Calculate $K_S$ and $K_A$*. If the second option (*Remove gaps and inframe stop codons*) is selected and the input is a pair of sequences, this option also allows to calculate $K_S$ and $K_A$ values by the codeml program (8) included in the PAML package. The calculation of $K_S$ and $K_A$ is only performed for sequence pairs because the computationally demanding construction of phylogenetic topologies would be required for alignments with more than two sequences. (iv) *Remove mismatches*. If there are mismatched codons between the protein and the DNA sequences, the users can either remove or retain such codon sites by this option. (v) *Use only selected positions*. With this option, only the codon alignment corresponding to the regions marked by '#' in the input alignment is generated. This option is very useful because it allows the construction of codon alignment for a certain exon or a domain or conserved blocks, for example those identified automatically by Gblocks (9). (vi) *Output format*. There are three output formats: CLUSTAL, PAML and FASTA. The output can be modified by combining these options (Figure 1D).

## CONCLUSION

The PAL2NAL server (http://www.bork.embl.de/pal2nal) is useful for constructing codon multiple alignments, which are required in many molecular evolutionary analyses, such as the calculation of $K_S$ and $K_A$ values. For a large-scale analysis, the distribution version of the PAL2NAL script, which is written in Perl and works in command line, is also available for download. We successfully applied the distribution version of PAL2NAL, for example, to the detection of human pseudogenes (10) and to the annotation of genes in human chromosomes 2 and 4 (11).

## ACKNOWLEDGEMENTS

The authors thank Yan P. Yuan for his help to set up the server and Eoghan Harrington for manuscript revision. The authors also thank Ziheng Yang for the permission to use codeml in the server and Formijn J. van Hemert for testing the server and providing us with valuable feedback. This work was supported by EU grant (LSHG-CT-2003-503265 and LSHG-CT-2003-503329). Funding to pay the Open Access publication charges for this article was provided by EMBL.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Miyata,T. and Yasunaga,T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino

acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.*, **16**, 23–36.
2. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
3. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
4. Wernersson,R. and Pedersen,A.G. (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.*, **31**, 3537–3539.
5. Bininda-Emonds,O.R.P. (2005) transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics*, **6**, 156.
6. Birney,E. and Durbin,R. (1997) Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 56–64.
7. 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.
8. Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.
9. Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
10. Torrents,D., Suyama,M., Zdobnov,E. and Bork,P. (2003) A genome-wide survey of human pseudogenes. *Genome Res.*, **13**, 2559–2567.
11. Hillier,L.W., Graves,T.A., Fulton,R.S., Fulton,L.A., Pepin,K.H., Minx,P., Wagner-McPherson,C., Layman,D., Wylie,K., Sekhon,M. *et al.* (2005) Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature*, **434**, 724–731.