

[IS-Online-Publication No. 16 | Berlin, July 2006]

align_dist

**a computer program for
pairwise alignment of DNA
sequences and
pairwise distances.**

INGO SCHINDLER, Berlin

Part I

File: align_dist.exe
for Windows

Part II

UserNotes

PART II: USER NOTES

DISCLAIMER

This VB-program is provided without any explicit or implicit warranty of correct functioning. To use this program on your computer is your own risk. The program has been developed for my own research program. Researchers can use this program for scientific and research purposes, but intellectual property and copyright for the source code and program remains the property of Ingo Schindler.

Contact:

Ingo Schindler, 12051 Berlin.

ingoschindler@web.de

1. Description

align_dist is a visual basic program for windows. It comprises two separate applications: (a) pairwise sequence alignment and (b) calculation of pairwise distances using different models of nucleotide substitution. The pairwise sequences alignments are stored in an output file which is used for the calculation for the distance matrices. It is also possible to calculate the distance matrix (bootstrap option available) of any particular alignment (e. g. a multiple alignment). However, it must be in the necessary data format or in **PHYLIP** (Felsenstein 2004) interleaved format.

There are much more sophisticated algorithms for calculating multiple alignments which are implemented in other free and more comfortable programs (e. g. T-Rex, MEGA, ClustalX). I wrote this small program to learn about the techniques in DNA-alignment and to keep the procedure simple as possible. I am convinced that it is very subjective to select an algorithm which reflect best the conceptions of the predicted results of the authors' point of view. The same is true for the subjective exclusion of regions of the DNA sequence or the use or not use of different codon positions. This is why I believe the procedure has to be simple as possible to fulfil the rules of an unbiased "parsimony".

2. Input file

a) Alignment

The input file must be in FastA format (see below for an example) in ASCII (*.txt, *.fas or fasta). No proceeding lines are allowed. The first letter must be a ">" (see 'Test Run' for further information).

b) Distance

The format for this application seems complicated in the first glance (see below). However, the application 'alignment' produce the data format during the procedure by itself, furthermore it can read PHYLIP format files (e. g. multiple alignment calculated by e. g. ClustalX). In such case (PHYLIP format file) there must be a proceeding line with "#phylip" instead of "#align".

3. Procedure

- a) Press 'Option' to select your preferred option for the calculation of the alignment and the distances.
- b) Select input file (use drive-, directory- and file-box). The input file format is tested by the program. If it is obviously wrong then the program does not open the file. In this case you should check the input file format.
- c) If the input file is a DNA sequence than press 'alignment' for starting the calculation of the pairwise alignments. If it is in a format for 'distance' than press 'distance' to start the calculation.

4. Test run

The test data are artificial without any biological meaning. The notes in brackets are only for explanation and are not part of the data file.

Input file (for alignment):

```
[fastA format; the first letter of every sequence must be a ">"; there are no proceeding rows allowed ]
>speciesA1 cytochrome b gene, partial sequence; mitochondrial [new line for the sequence]
ACGGTAGTTGAGPTTA [new line for the next sequence]
>speciesB2 cytochrome b gene, partial sequence; mitochondrial
ACTAAGGTAGTGTAATA [use only capital letter ACTG]
>speciesC3 cytochrome b gene, partial sequence; mitochondrial
ACGGTAGTTGAGTTA
>speciesD4 cytochrome b gene, partial sequence; mitochondrial
ACTAAGGTAGTGTAATA
```

Result and Input (input format for the calculation of the distance matrix):

```
#align [the input file for the 'distance' must be start with "#align".]
  4 [number of sequences]
speciesA1_ [name of sequence / species]
speciesB2_
speciesC3_
speciesD4_
1 [sequence 1]
2 [sequence 2]
>
```

ACGGTAGTTGAGAT-TAATTG [alignment sequence 1 vs sequence 2]

<

>

ACTA-AGGTAGTGTAATAATT [use only capital letter ACTG or "-"]

<

1

3

>

ACGGTAGTTGAGATTAATTG [alignment sequence 1 vs sequence 3]

<

>

ACGGTAGTTGAG-TTAAGTC

<

1

4

>

ACGGTAGTTGAGATTAATTG

<

>

ACTA-AGGTAGTGTAAGC-

<

2

3

>

ACTAAGGTAGTGTAATAATT

<

>

ACGGTAGTTGAGTTAAG-TC

<

2

4

>

ACTAAGGTAGTGTAATAATT

<

>

ACTAAGGTAGTGTAAG--C

<

3

4

>

ACGGTAGTTGAGT-TAAGTC

<

>

ACTA-AGGT-AGTGTAAGC

<

Date Time (start)
 Time (finish)

Scores [the scores used for the alignment calculation]

match = 2

'transition' = 1

```
'transversion' ==-1
gap penalty ==-2
```

Output

```
-----
Date      Time
-----
Distanz Matrix
  _____JUKES & CANTOR_____
  4
speciesA1_ 0 .7489 .1135 1.0126
speciesB2_ .7489 0 .7489 .1202
speciesC3_ .1135 .7489 0 .3733
speciesD4_ 1.0126 .1202 .3733 0

  _____Kimura 2P_____
  4
speciesA1_ 0 .7631 .1147 1.0986
speciesB2_ .7631 0 .7631 .1257
speciesC3_ .1147 .7631 0 .3741
speciesD4_ 1.0986 .1257 .3741 0
```

5. Options

Alignment: Dynamic programming algorithm (Needleman & Wunsch 1970) is used to create the pairwise alignment as it is described e.g. in Cannarozzi (2005, Graur & Li 1999). Using this procedure a scoring matrix for scoring substitutions, matches and gap creating is needed. You can change every single parameter. However, not every combination or value makes sense. Therefore the appropriate scores (from Cannarozzi 2005) are summarized in Table 1. For an introduction of molecular evolution see e.g. Graur & Li (1999). There are also text books in German which are valuable for overview and introduction, e.g. Böckenhauer & Bongartz (2003), Hütt & Dehnert (2006).

Tab.1

Default values for the scoring Matrix.
Gap penalty -2.

	A	C	G	T
A	2	-1	1	-1
C	-1	2	-1	1
G	1	-1	2	-1
T	-1	1	-1	2

Distance: There are several methods for estimating the number of nucleotide substitutions (e. g. Gojobori et al. 1990). Within this program three different methods are available. The *One-Parameter Method* (also known as the Jukes & Cantor corrector) using the equation (see e. g. Gojobori et al. 1990):

$$(1) \quad d_{ij} = 0.75 \log(1 - \frac{4}{3} P),$$

where P is the observed proportion of different nucleotides between the two strings (Graur & Li 2000). For the Two-Parameter Method (also known as the Kimura-2P model) the difference between two sequences is divided into transitions and transversions. For this substitutions model the equation (see e. g. Gojobori et al. 1990)

$$(2) \quad d_{ij} = -0.5 \log(1 - 2P - Q) - 0.25 \log(1 - 2Q)$$

is used. P and Q are the proportions of transitional and transversional differences between the two strings, respectively (Graur & Li 2000). Among different variation of the LogDet methods the program used the formula (adopted from Tamura & Kumar 2002):

$$(3) \quad d_{ij} = -0.25 \log(\det F_{ij}) - \log(4),$$

where the function *det* stands for the determinant of the F_{ij} matrix (Tab. 2).

Tab. 2 (adopted from Wägele 2000)

$$F_{ij} = \begin{vmatrix} f_{AA} & f_{AC} & f_{AG} & f_{AT} \\ f_{CA} & f_{CC} & f_{CG} & f_{CT} \\ f_{GA} & f_{GC} & f_{GG} & f_{GT} \\ f_{TA} & f_{TC} & f_{TG} & f_{TT} \end{vmatrix}$$

Bootstrap: Select the number of bootstrap runs. Be aware that the first matrix is calculated on the unmodified strings. That's mean that if you select 100 you get 101 distance matrices of the selected substitute model.

6. Notes

- 1) The program saves the results in the file with the name 'alignX' or 'distanceX' respectively. X = is a running number. The program does not check if this file name already exists. Thus, it may happen that your old data file is deleted.

- 2) If the input data file format is not correct the results can be wrong or the program terminates without further information or warning.
- 3) The limitation for the length of the sequence is bp 1200 and the limit of the number of sequences is 50.
- 4) The program creates a 'StartPfadAD.ini' file, which contain the start path and parameter used in the last running.
- 5) Be careful with large data sets, because the program is slow.
- 6) Although I tested the results many times, there is no guarantee that the results are correct. To use this program is your own risk.
- 7) If you need only ordinary pairwise distances of a multiple alignment than use much faster and more comfortable programs like MEGA, PHYLIP or T-REX (which are also free).
- 8) If you need only a simple multiple alignment than use e. g. ClstalX, MEGA or T-REX.

7. Literature

Böckenhauer, H.-J. and D. Bongartz (2003): Algorithmische Grundlagen der Bioinformatik – Modelle, Methoden und Komplexität. Teubner, Stuttgart.

Cannarozzi, G. M. (2005): String alignment using dynamic programming.
<http://biorecipes.com/DynProgBasic/code.html> [assess 06. VI 2006]

Felsenstein, J. (2004) PHYLIP: the phylogenetic inference package.
<http://evolution.genetics.washington.edu/phylip.html> [assess 06. VI 2006]

Felsenstein, J. (2004a): Inferring phylogenies. Sinauer, Sunderland, Massachusetts.

Gojobori, T., E. N. Moriyama and M. Kimura (1990): Statistical methods for estimating sequence divergence. Methods in Enzymology 183: 531-550.

Graur, D. and W.-H. Li (2000): Fundamentals of molecular evolution. 2nd edition. Sinauer, Sunderland, Massachusetts.

Hütt, M.-T. and M. Dehnert (2006): Methoden der Bioinformatik – Eine Einführung. Springer, Berlin.

Needleman, S. B. and C. D. Wunsch (1970): A general method applicable to search for similarities in the amino acid sequence of to proteins. J. Mol. Biol. 48: 443-453.

Tamura, K. and S. Kumar (2002): Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol. Biol. Evol. 19(10):1727–1736.

Wägele, J.-W. (2000): Grundlagen der phylogenetischen Systematik. Verlag Dr. Friedrich Pfeil, München.

If you find bugs or mistakes in this program, please contact me immediately!
Any help and suggestions are welcome.

Please cite:

Schindler, I (2006): *align_dist*, a computer program for pairwise alignment of DNA sequences and pairwise distances.

IS-Online-Public. No. 16.