

ClustalG: Software for analysis of activities and sequential events

Clarke Wilson
Canada Mortgage and Housing Corporation
C5-318 700 Montreal Road
Ottawa, Canada K1A 0P7
Phone 613 748 4670
Fax 613 748 4865
email cwilson@cmhc-schl.gc.ca

Andrew Harvey
Department of Economics
St. Mary's University
Halifax, Nova Scotia B3H 3C3
Phone 902 420 5676
email andrew.harvey@stmarys.ca

Julie Thompson
Intitute de Genetique et de Biologie Moleculaire et Cellulaire
Strasbourg, France
julie@igbmc.u-strasbg.fr

1999

Paper presented at the Workshop on Longitudinal Research in Social Science: A Canadian Focus, Windermere Manor, London, Ontario, Canada, October 25-27, 1999.

ABSTRACT

The paper describes a new software package for sequence alignment analysis called ClustalG. The package is a rewrite of the well-known Clustal series of alignment packages. The main new feature of ClustalG is the recognition of input word sequences of up to six characters. This effectively eliminates the 20 letter constraint implemented in biological software on the number of event categories available to the researcher.

The essential ClustalG windows are shown and the meaning of the most common variable settings are discussed. Some elementary alignments and clustering trees are illustrated. The software package including the help file and a number of time use sample files is freely available to any user from the St. Mary's University ftp site.

Key words: activity patterns sequence alignment software

Acknowledgements

The authors acknowledge the original work on the Clustal program group by Des Higgins and Paul Sharp, and we thank Toby Gibson of the European Molecular Biology Laboratory for agreeing to allow us to redesign the program for use outside molecular biological applications. Their work represents a huge investment of research funds and talent that will now be applied to subjects far beyond those for which Clustal was originally designed.

This work is part of the *Activity Settings, Sequencing and the Measurement of Time Allocation Patterns*, project funded by the Social Science and Humanities Research Council of Canada.

1. The ClustalG project

The *Activity Settings: Design, Measurement, and Analysis* research project funded by the Social Science and Humanities Research Council of Canada in 1994 produced a number of papers and publications that have illustrated the application of sequence alignment methods and software as developed in molecular biology to time use and transportation research [1, 2, 3]. These have all used versions of the Clustal programs maintained at the European Molecular Biology Laboratory. The results of the applications suggest that alignment methods hold great promise for examining social processes that consist of sequences of activities. Abbott [4] has reviewed a variety of research into sequential processes based on alignment or optimal matching as the methods are sometimes called. However, the biological software contains a number of features that have no place in social science research, and available packages generally limit the eligible alphabet to just over 20 characters.

A subsequent SSHRCC project, *Activity Settings, Sequencing and the Measurement of Time Allocation Patterns*, has contracted the Clustal programmer, Julie Thompson, to amend the windows version, ClustalX, for the research in any discipline that deals with sequential processes.

The product is called ClustalG (for general) and is available from the ftp site at St. Mary's University, Halifax.

The properties of ClustalW and ClustalX have been published [5,6]. Briefly, the packages implement a two stage process of calculating the pairwise similarities in a set of sequences then constructing a tree from transformations of the similarities. The tree is used to guide the progressive multiple alignment of the set of sequences.

ClustalG has deleted the explicitly biochemical features of ClustalX, has expanded the input routines to accept multiple letter words of up to six characters, and has created a new output file that specifies the members of each step by which the program clusters individuals into progressively larger and more general pattern groupings. The key feature is the introduction of multiple letter words because this permits analysts to use complex coding schemes that are usual in many sciences. Analysts may use different positions in the word to indicate different dimensions of events. In our example data, the first two positions indicate an activity, the third indicates location, and the fourth who else was present.

2. Sequence alignment methods

Sequence alignment, or optimal string matching as the methods are also called, employ combinatorial algorithms to calculate measures of either similarity or distance between character sequences. See Waterman [7] for a comprehensive treatment of alignment mathematics and biological applications. When stages of processes or activities are represented by characters, these measurements can form the basis of taxonomies of the behaviour being examined. Alignment methods provide the most rigorous basis available for classifying groups of character sequences.

The general process can be illustrated by writing the elements of two sequences in the margins of a comparison table and placing an asterisk in cells for which marginal elements match. Consider the comparison of letters of [mississippi] and [missouri] shown in Figure 1.

Figure 1: Comparison table for [mississippi] and [missouri]

	<u>m</u>	<u>i</u>	<u>s</u>	<u>s</u>	<u>i</u>	<u>s</u>	<u>s</u>	<u>i</u>	<u>p</u>	<u>p</u>	<u>i</u>
m	*										
i		*			*			*			*
s			*	*		*	*				
s			*	*		*	*				
o											
u											
r											
i		*			*			*			*

The degree similarity of the two names is established in the first syllable as shown by the downward sloping diagonal pattern of stars. The [iss] substring is repeated in [mississippi] and

this is illustrated by the second diagonal, translated three positions to the right. The remaining letter matches are more or less random.

The alignment algorithms are based on calculation of a cumulative score beginning at the upper left cell and proceeding to the lower right. A cell's score is based on the preceding score plus its own value. Values are determined by weighting systems related to the substance of the problem in question. A path can be found that leads backwards from the lower right cell through the highest value cells to the upper left. The order in which letters are included in the path, and in particular whether a letter matches another letter or is placed against a gap, determines the pairwise alignment. Gaps may be inserted in either sequence to allow identical letters to match. Optimal paths and alignments are often not unique. One option for the alignment of [mississippi] and [missouri] is shown below:

```

m i s s - - - i s s i p p i
m i s s o u r i - - - - -

```

The exact patten of letters in positions five and following is determined by the system of scoring weights and gap penalties used.

Pairwise alignment may be generalized to multiple alignments by defining comparison tables and paths in N dimensions. However, for N greater than about 10 sequences, the algorithms are prohibitively costly in time and memory space. Multivariate alignments are usually implemented using approximate methods based on pairwise measures. This is the case with the Clustal program family.

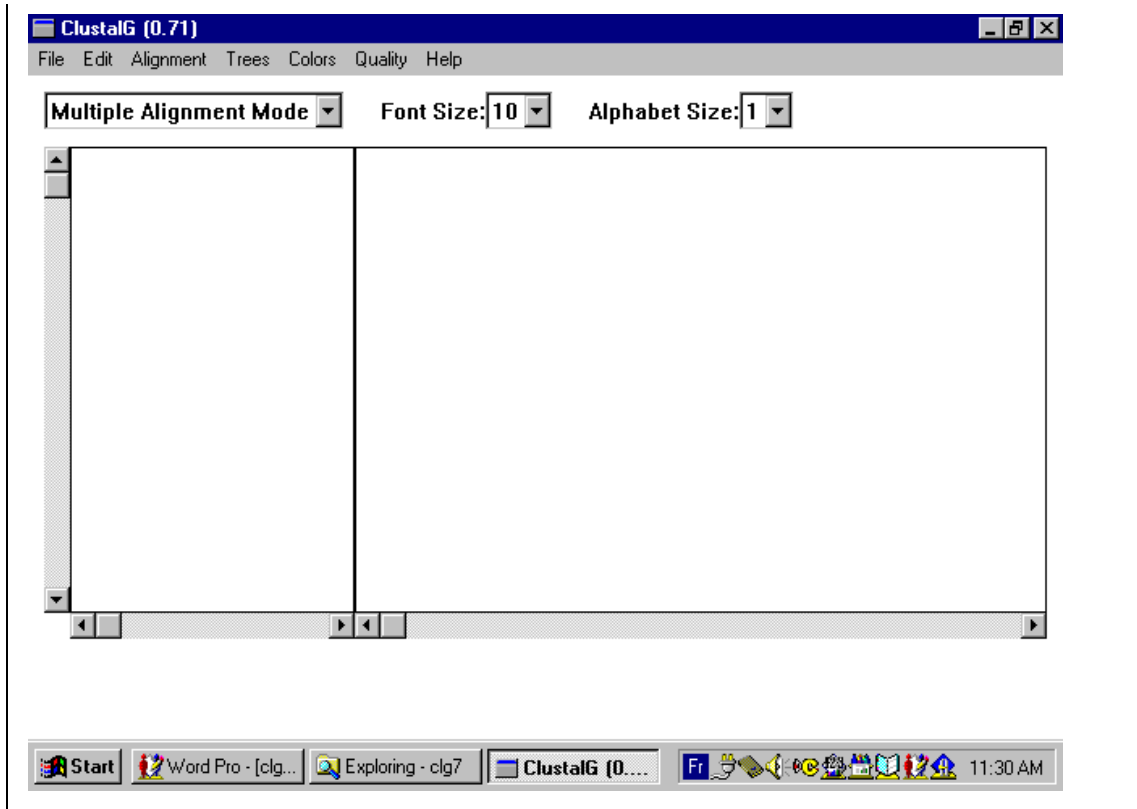
3. ClustalG screen

The ClustalG screen is shown in Figure 1 as it is displayed when the program is executed. Seven menus control the loading of sequence files, editing, alignments, preparation of trees calculated as a result of the clustering process performed progressively on the sequence file, coloring, depiction of special sequences or segments (quality), and the help screens. This presentation deals only with file manipulation and specification of alignment parameters. The ClustalG online help facility covers the other items.

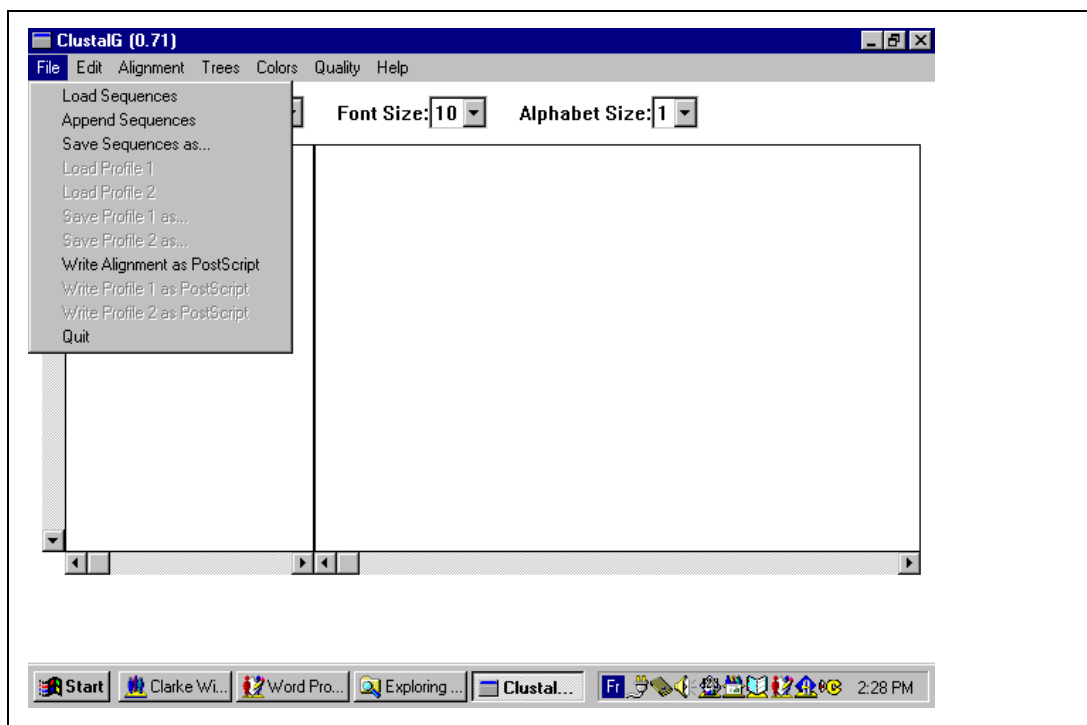
ClustalG operates in Multiple Alignment or Profile Alignment modes, which are selected from the first of the drop down boxes. Multiple Alignment mode uses a single screen. Profile Alignment mode uses two screens because a profile alignment is an alignment of two previously constructed alignments. Multiple alignment mode is normally used first to find useful arrangements of sequence data. The researcher may later want to combine various alignments.

The Alphabet Size from one to six characters must be chosen before sequences are loaded in to ClustalG. All elements of all sequences are treated as having a constant size.

The Windows menu bar at the bottom is not part of the ClustalG screen.



4. File menu options



Load sequences:

This is the first step and is mandatory. Selection of the load sequence option invokes a Windows *Open* screen that allows user to specify drive, folder and filename. Sequence labels are written in the left-hand box and the sequence elements are written to the right.

Many single letter sequence formats are used in biology. ClustalG allows all that have been implemented in ClustalX in addition to the multiple letter words. The simplest is the Pearson or Fasta format which has been used in the example files. This format begins each sequence record with a greater-than symbol and the characters on the line following are treated as a label. Lines following the first line are treated as sequence data and are read until another greater-than symbol or the end of file character is found. For example:

```
> 1346wda 12e 12
rewaeawreamr
> 1444sna 17e 17
rrkcdedcacedmckkmr
```

An example of a sequence that uses 4 letter words is:

```
> 2011mna 15e 15
ZzhaPchaPchaTrtaWkwaWkwaWkwaTrtaZzhaTvhaEthfZzhaZzhaFchfZzha
```

The diary reads: asleep, home, alone; personal care, home, alone; personal care, home, alone; travel, location is travel, alone; work, at workplace, alone; etc...

Append sequences:

Additional sequences can be added to a file previously loaded.

Save sequences as:

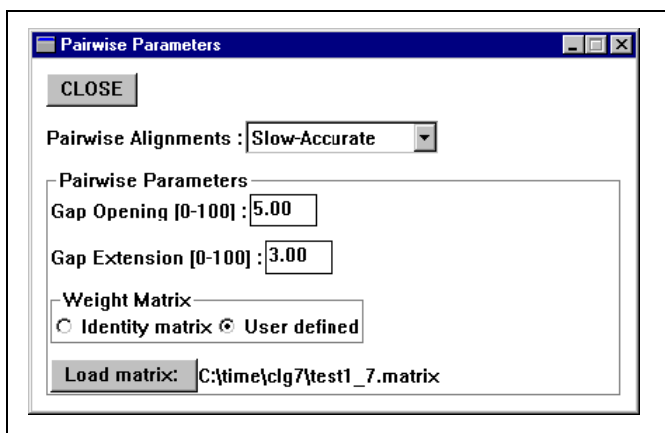
Permits user to specify a new file location for edited sequence files or for new alignments using different sets of parameter values.

Profile options:

Similar to the multiple alignment options except that two files are specified

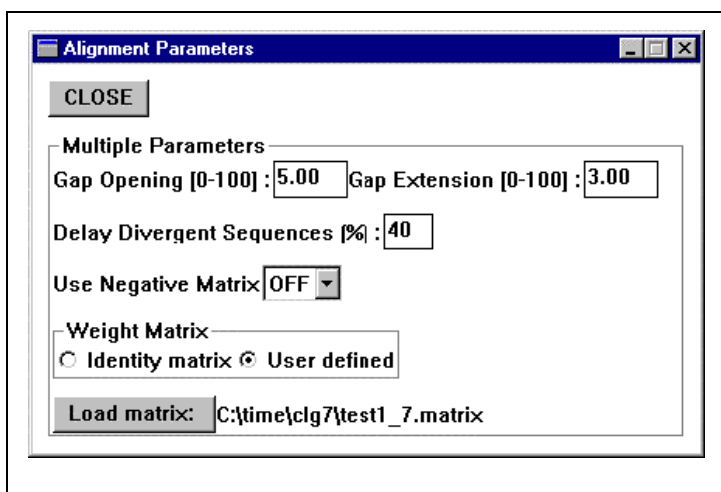
Write as Postscript:

Creates Postscript graphic output file.



Multiple Alignment Parameter screen

This options invokes another dialogue box the allows the user to control the set of parameters used by ClustalG in conjunction with pairwise similarity scores to calculate a guide tree, and from there to assemble the multiple alignment. The weight matrix usage is the same as for the pairwise screen.



Gap parameter option screen

A dialogue box allows further specification of gap parameters.

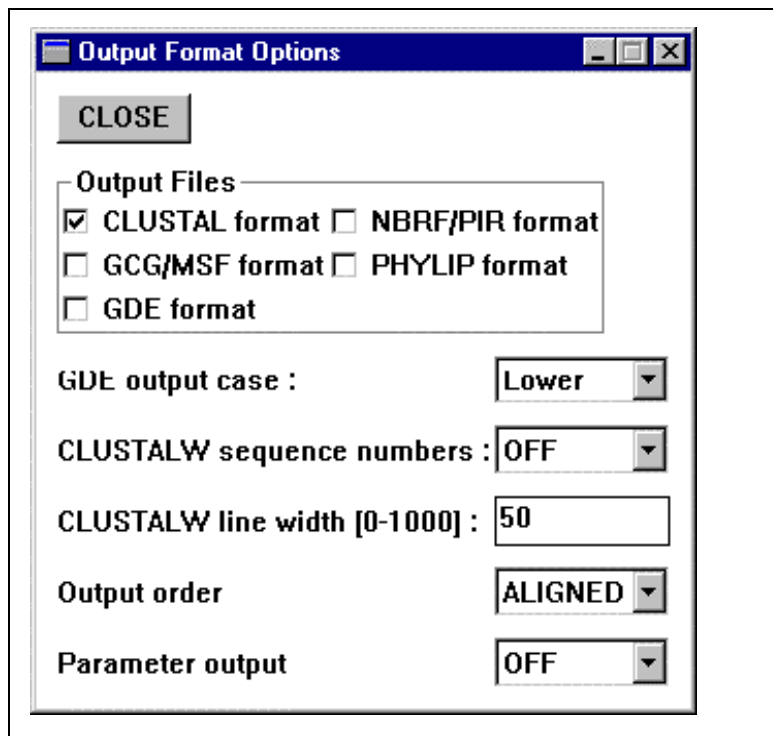
Output format option screen

User selects one or more formats for the output files

The line width may be set to control alignment appearance. The alignment is written as a series of blocks of fixed width. Where output lines are comparatively short, they may fit on letter or legal paper in portrait or landscape orientation. Where lines are too long the user can control block width up to 1000 characters.

The aligned sequences may be written in input order or in an order that roughly follows their grouping order.

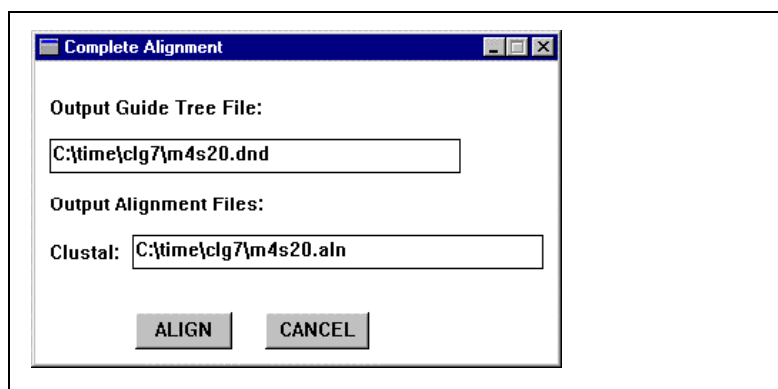
The user may parameter set future



choose to have the written to a file for reference.

Do Complete Alignment Screen

This screen allows the user to name the output alignment and dendrogram files. The default is to use the same name as the sequence file



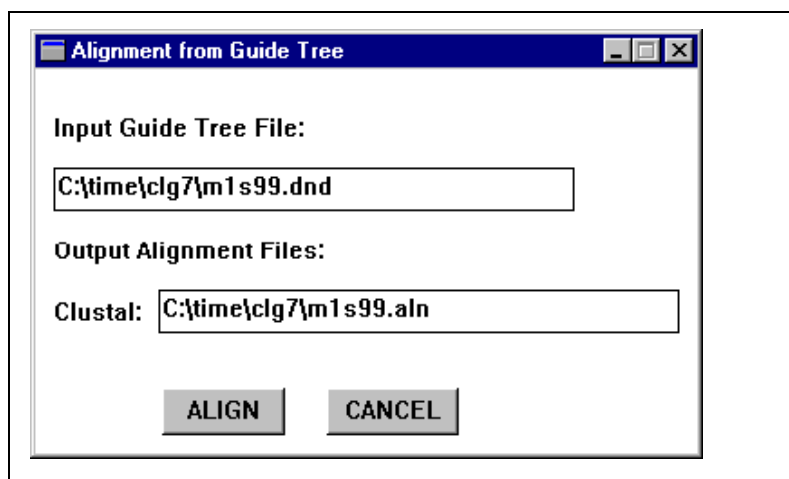
that was loaded with extensions of *.aln and *.dnd. An alignment file is shown later. The guide tree or dendrogram is written as a text file of nested parentheses containing sequence labels and the branch lengths of the tree which can be drawn from the nesting pattern. No tree is drawn. However, the file format is recognized by biological graphics software.

Produce guide tree only

This generates only the *.dnd file. This may be used with tree drawing software (for example Treeview by Rod Page, University of Glasgow, or Phylip by Joe Felsenstein , University of Washington) to display the dendrogram graphically. A Treeview [8] screen is shown later.

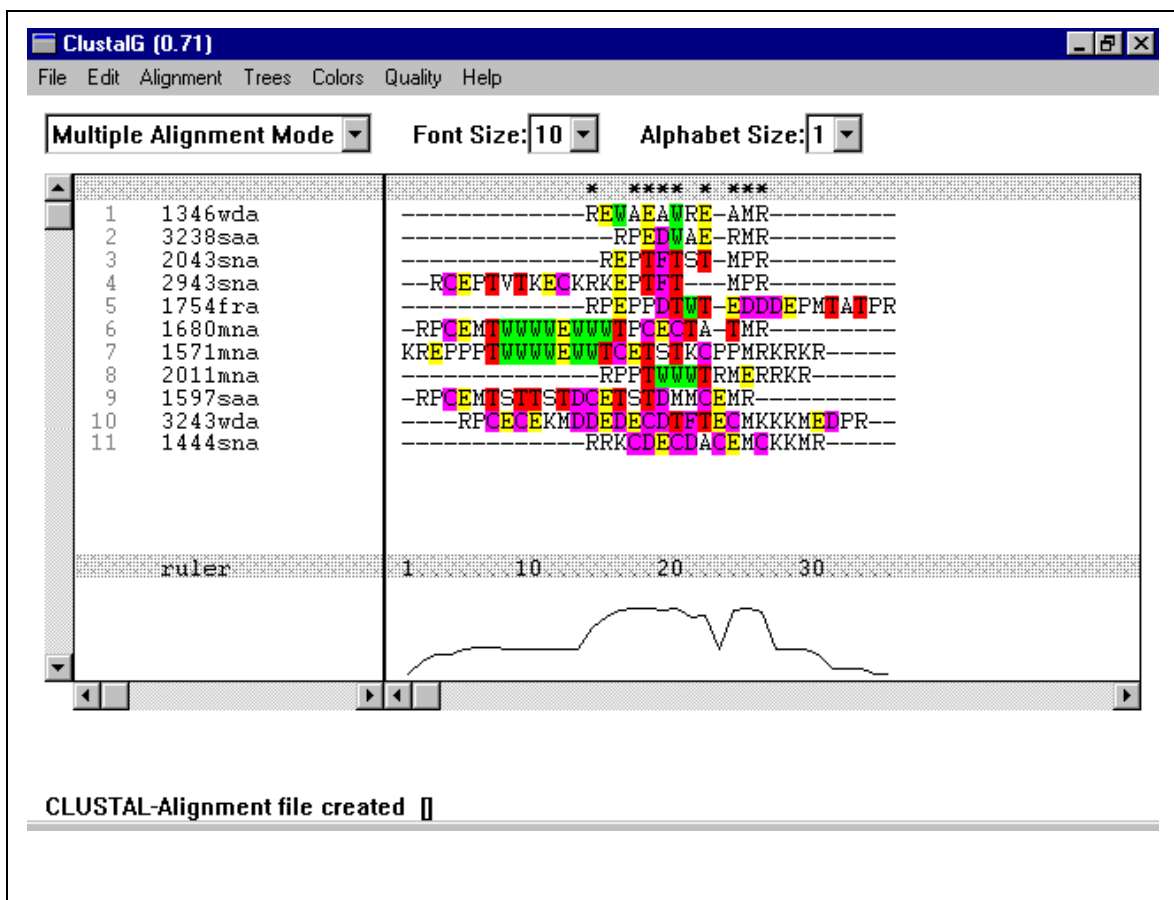
Alignment from guide tree screen

Specifies an existing guide tree to control the multiple alignment. New multiple parameters may be selected.



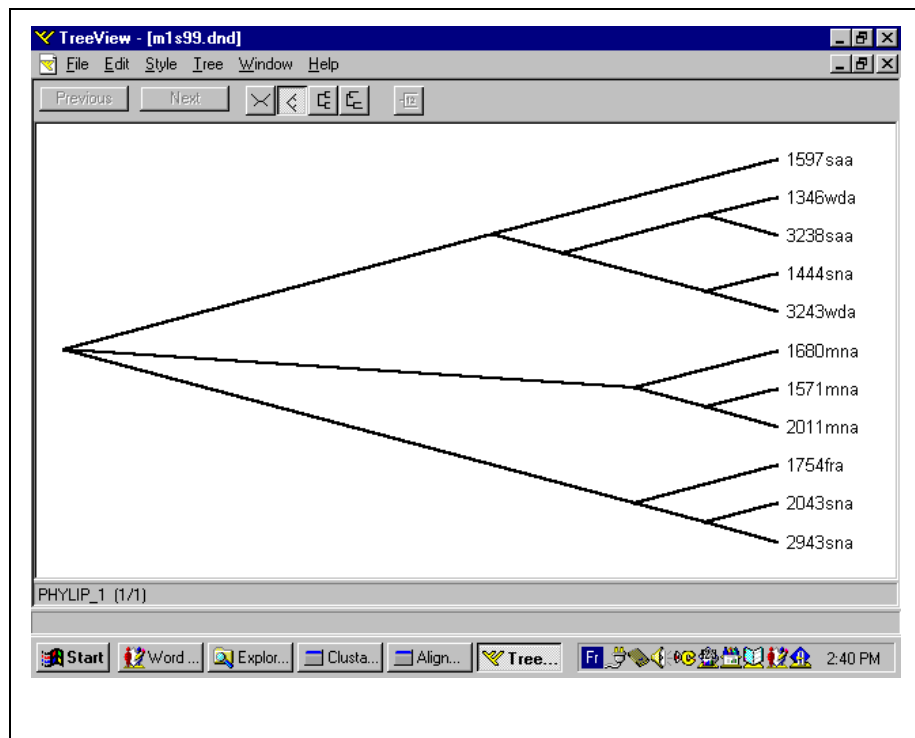
6. ClustalG Alignment Screen

ClustalG has identified three primary behavioural groups in the test data file. Their multiple alignment is shown above and the groupings are shown in the tree diagram on the next page. The order of output in the alignment is roughly but not precisely that in the tree diagram and the two should be used together to identify the membership of the primary behavioural groupings that occur in the alignment.



7. Treeview illustration of the ClustalG guide tree file

The tree identifies similar groups of sequences precisely. The alignment describes what the activity patterns are. The middle group in the tree diagram are employed people who had several work episodes in their diaries.



References

1. Wilson W.C. 1998 Activity pattern analysis by means of sequence alignment methods, *Environment and Planning*, volume 30, pp. 1017-1038
2. Wilson W.C. 1998 Analysis of travel behaviour using sequence alignment methods, *Transportation Research Record*, number 1645, pp. 52-59.
3. Harvey A.S. and Wilson W.C. 1998, Evolution of daily activity patterns: a study of the Halifax panel survey, paper presented at Thematic Group 1, Time-Use, World Congress of Sociology (in conjunction with Association, International Association for Time Use Research) University of Quebec, Montreal, July 26-August 1, 1998.
4. Abbott A. 1999 the review paper. citation to come
5. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. 1997, The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, 24:4876-4882.
6. Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673-4680.
7. Waterman, M. 1995, *Introduction to Computational Biology*, Chapman and Hall, London
8. Page, R. D. M. TREEVIEW: An application to display phylogenetic trees on personal computers, *Computer Applications in the Biosciences*, 12:357-358.