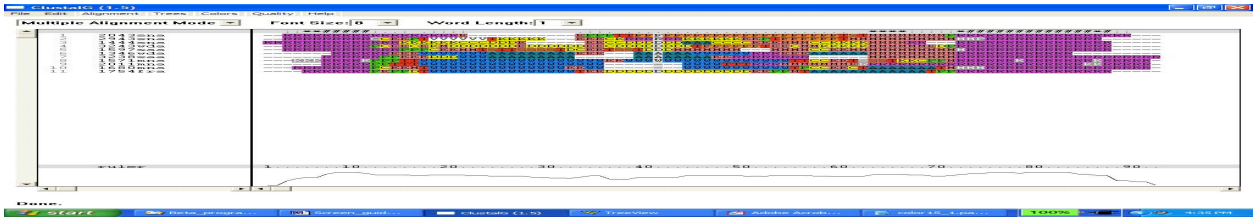


DRAFT

ClustalG: Software for analysis of activities and sequential events

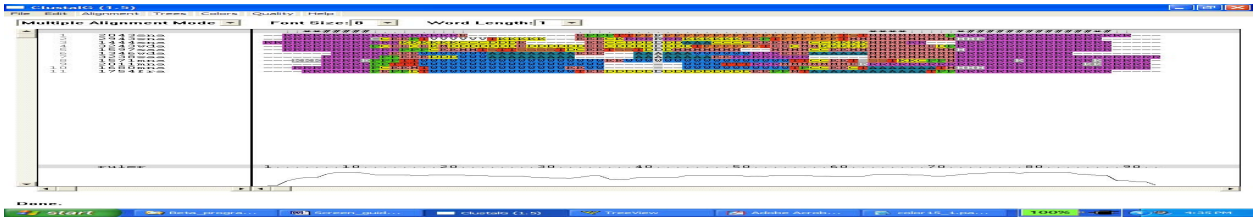
Clarke Wilson
Canada Mortgage and Housing Corporation
C5-318
700 Montreal Road
Ottawa, Canada
K1A 0P7

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
Institut de Genetique et de Biologie Moleculaire et Cellulaire
Strasboug, France
julie@igbmc.u-strasbg.fr



DRAFT

ABSTRACT

The paper describes a software package for sequence alignment analysis called ClustalG. The package is a rewrite of the well-known Clustal series of alignment packages written for protein and nucleic acid analysis. The main feature of ClustalG that distinguishes it from the versions used by biochemists is the recognition of input word sequences of up to 12 characters. This effectively eliminates the 26-letter constraint implemented in biological software on the number of event categories available to the researcher.

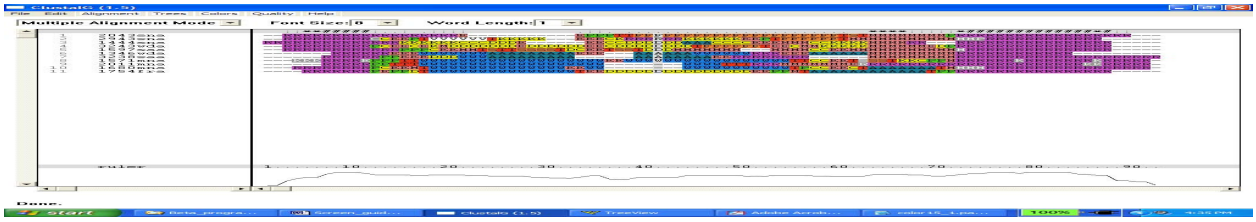
The essential ClustalG windows are shown and the meanings of the most common variable settings are discussed. Some elementary alignments and clustering trees are illustrated. The software package including the help file and time use sample files is freely available to any user from the website of the International Association for Time Use Research at St. Mary's University: www.stmarys.ca/partners/iatur/clustalG.

Key words: time use, activity patterns, sequence alignment software

Acknowledgements

The authors acknowledge the original work on the Clustal program group by Des Higgins and we thank Toby Gibson of the European Molecular Biology Laboratory for agreeing to allow us to use program enhancements made at that laboratory as the basis for a generalized version 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.

The production of ClustalG was part of the *Activity Settings, Sequencing and the Measurement of Time Allocation Patterns*, project directed by Andy Harvey and funded by the Social Science and Humanities Research Council of Canada.



DRAFT

1. The ClustalG project

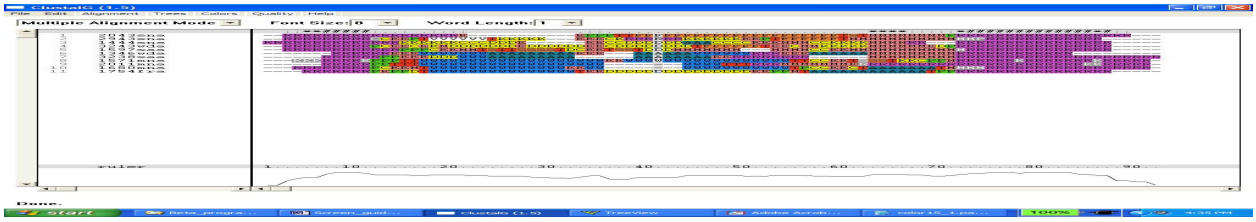
The *Activity Settings: Design, Measurement, and Analysis* research project funded by the Social Science and Humanities Research Council of Canada in 1994 and directed by Andrew Harvey produced a number of papers and publications that have illustrated the application of sequence alignment methods and software as developed in molecular biology to time use and transportation research [1, 2, 3]. These all used versions of the Clustal programs maintained at the European Molecular Biology Laboratory in Heidelburg. The results of the applications suggested that alignment methods hold great promise for examining social processes that consist of sequences of activities. Abbott [4] 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 26 characters which is sufficient to illustrate all of the amino acids that compose proteins and the nucleotides that comprise DNA.

A subsequent SSHRCC project, *Activity Settings, Sequencing and the Measurement of Time Allocation Patterns,* 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 freely available to any user from the website of the International Association for Time Use Research at St. Mary's University : www.stmarys.ca/partners/iatur/clustalG.

The properties of ClustalW and ClustalX are described in a number of publications [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 12 characters, and has created an 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.

The expansion of the elements of the input sequences to multi-letter words also forces the analyst to know more about the relative similarities of elements that may be identical on one or more dimensions, but may differ on others. The user passes these complex similarity relations to



ClustalG in parsing files that specify relative similarities of elements or words that differ in specific places.

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. Waterman [7] gives a comprehensive treatment of alignment mathematics and biological applications. When the stages of processes or activities are represented by characters alignment similarity measures can form the basis of taxonomies of the behaviour being examined. Alignment methods provide a rigorous basis for classifying groups of character sequences.

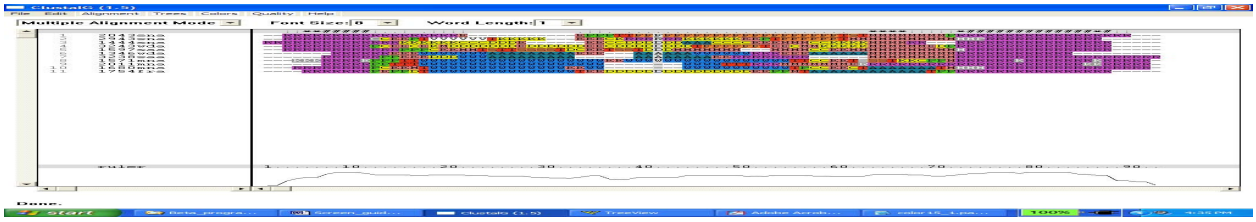
Writing the elements of two sequences in the margins of a comparison table and placing an asterisk in cells for which marginal elements match illustrates the general alignment process. Consider the comparison of letters of [mississippi] and [missouri] shown in Figure 1.

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

	m	i	s	s	i	s	s	i	p	p	i
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] sub-string 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 scoring 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 pattern 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.

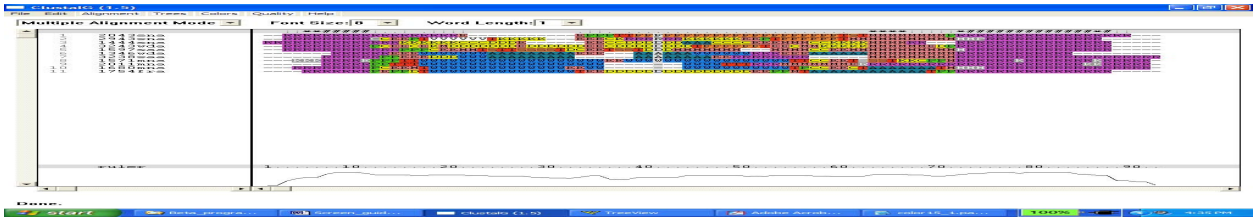
Many single letter sequence formats are used in biology. ClustalG allows all that have been implemented in ClustalX in addition to the multiple letter words. The simplest is the Pearson or Fasta format, which has been used in the example files. This format begins each sequence record with a greater-than 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. The following sequence is a daily activity diaries containing 12 events.

```
> 1346wda 12e 12
rewaeawreamr
```

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

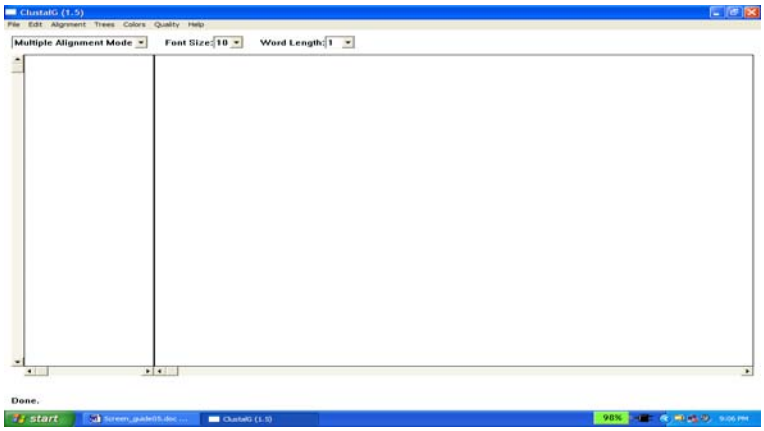
```
> 2011mna 15e 15
ZzhaPchaPchaTrtaWkwaWkwaWkwaTrtaZzhaTvhaEthfZzhaZzhaFchfZzha
```

The diary reads: asleep, home, alone; personal care, home, alone; personal care, home, alone; travel, location is travel, alone; work, at workplace, alone; etc. The Courier font is convenient because it spaces all characters equally.



DRAFT

ClustalG main screen

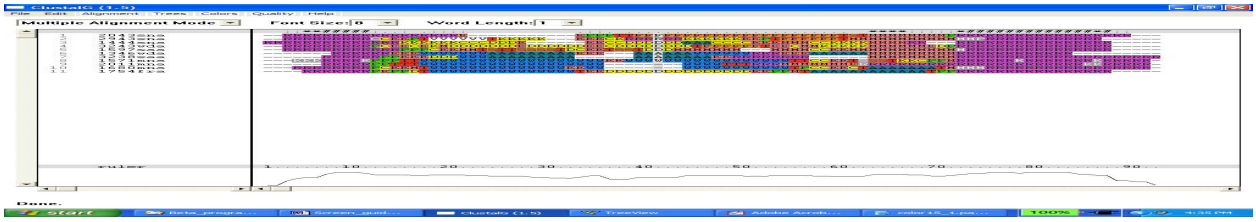


The ClustalG main screen is shown in Screen 1 as it is displayed when the program is executed. Seven menus control the loading of sequence files, editing, alignments, preparation of trees calculated as a result of the clustering process performed progressively on the sequence file, coloring, depiction of special sequences or segments (quality), and the help screens. This presentation does not deal with the Quality or Help menus. The Help menu gives the definitive account of all features.

The Windows menu bar at the bottom is not part of the ClustalG screen. Phrases in Arial 10 type refer to ClustalG options.

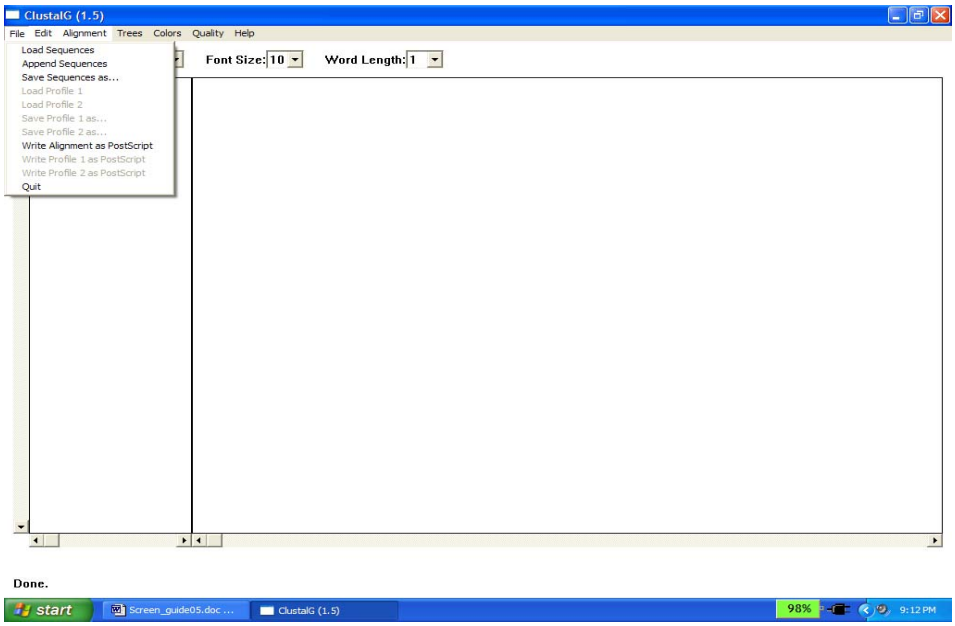
ClustalG operates in Multiple Alignment or Profile Alignment modes, which are selected from the first of the drop down boxes. Multiple Alignment mode uses a single screen. Profile Alignment mode uses two screens because a profile alignment is an alignment of two previously constructed alignments. The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

A Word Length from one to 12 characters must be chosen before sequences are loaded in to ClustalG. All elements of all sequences are treated as having a constant size.



DRAFT

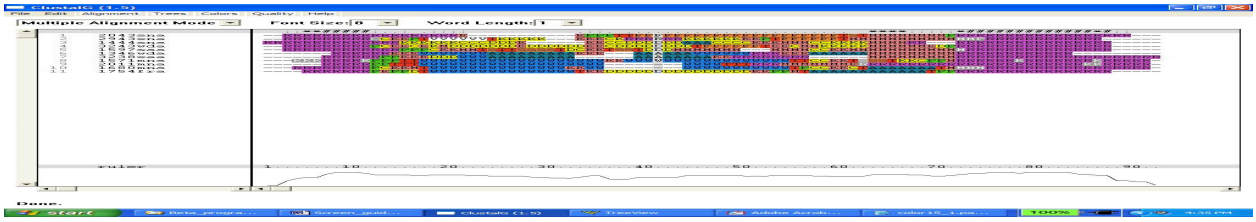
Screen 1 File menu options



In the multiple alignment mode, five options are available as shown in Screen 1.

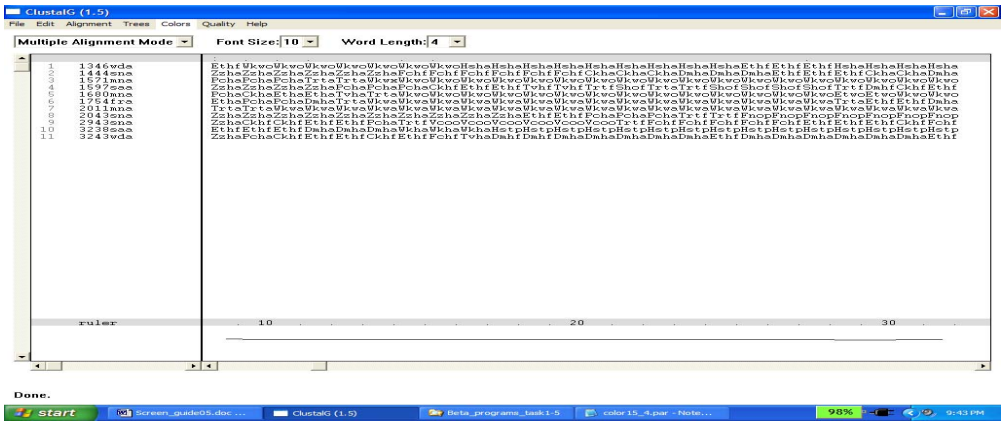
Load sequences: This is the first step and is mandatory. Selection of the load sequence option invokes a FileOpen screen that allows user to specify drive, folder and filename. Sequence labels are written in the left-hand box and the sequence elements are written to the right.

Screen 1a, below, shows the loaded sequences with a word length setting of four. Note that for readability, the first character of the word is in upper case and the remainder are in lower case. It is useful to create a name for the sequence file that describes the experiment being carried out, e.g. name.seq. ClustalG creates a number of output files with names that default to the sequence input name but which have distinctive extensions.



DRAFT

Screen 1a

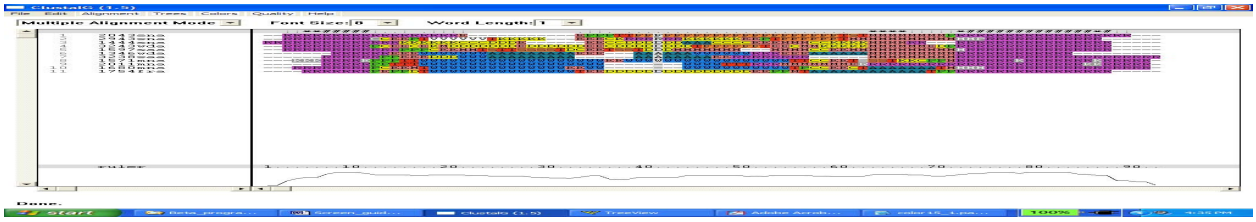


Append sequences: Additional sequences can be added to a file previously loaded.

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.

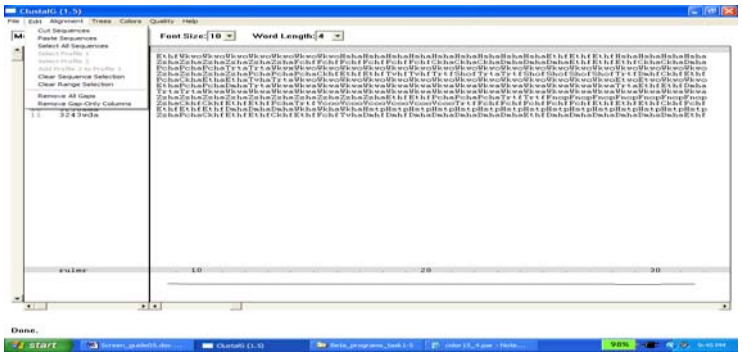
Profile options: Similar to the multiple alignment options except that two files are specified.

Write as Postscript: Creates Postscript graphic output file. These files can be read with Adobe Standard software and converted into *.pdf files. They are useful for preserving the coloring of sequence alignments.



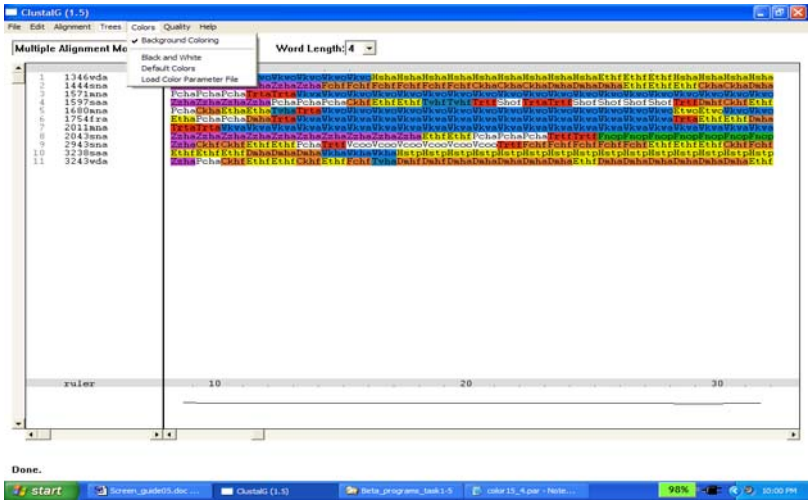
DRAFT

Screen 2 Edit options

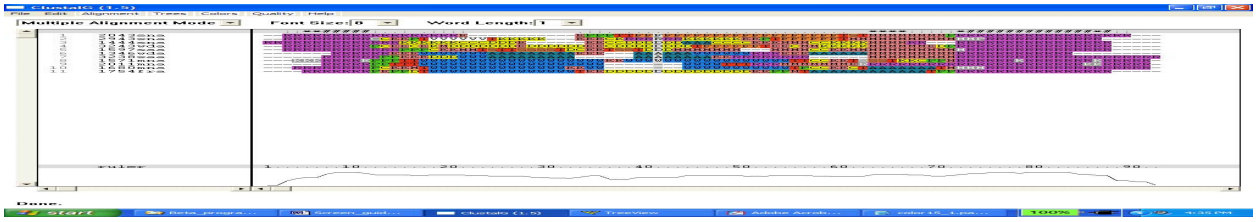


The edit options allow the user to add or remove sequences from a loaded file or to rearrange the appearance of an aligned file.

Screen 4 Color files

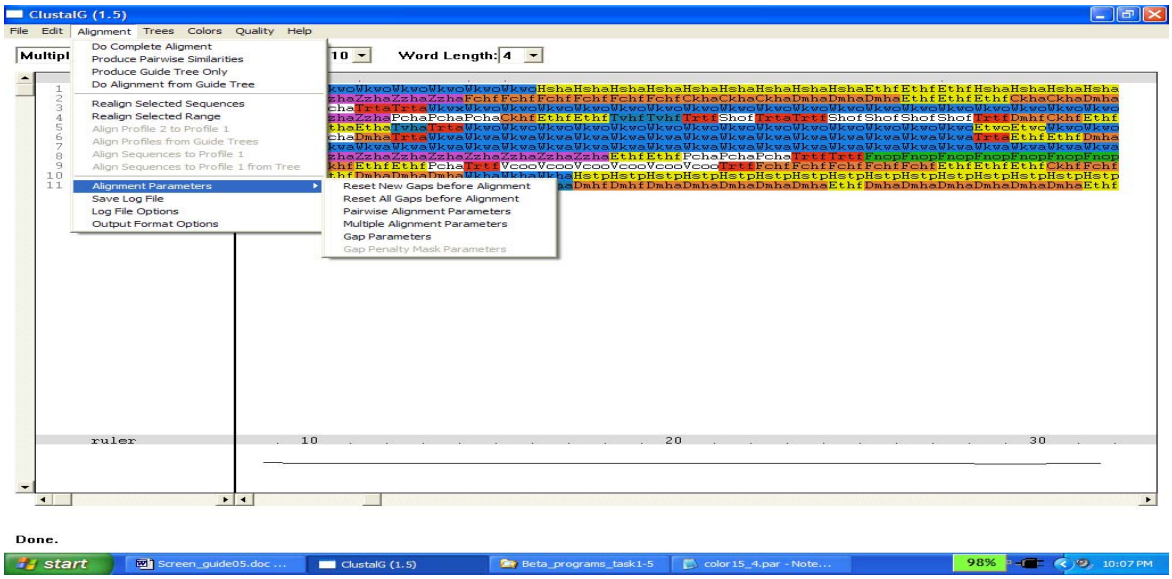


Background color	Colors may be associated with the background or lettering of the sequences. The example shows background coloring.
Load Color File	Selection invokes a file selection window that allows the user to choose a with color parameter file. The instructions for preparing color parameter files are available in the Help menu.



DRAFT

Screen 3 Alignment menu

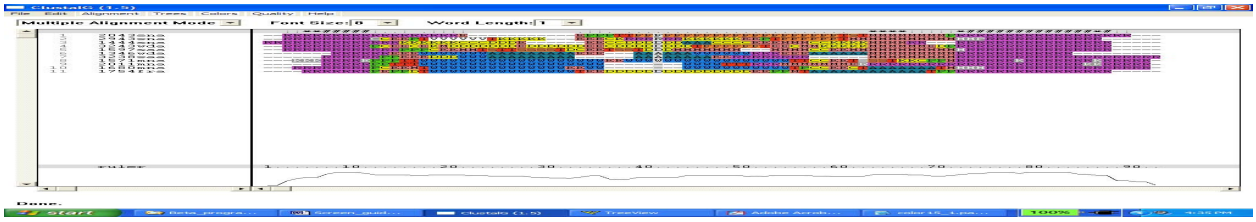


The first subgroup of options launches and controls the output of ClustalG files, including the alignment file (*.aln), the dendrogram file (*.dnd). The second group allows amendment of existing alignments, and the third group sets the parameters for the alignment. Group three is described first.

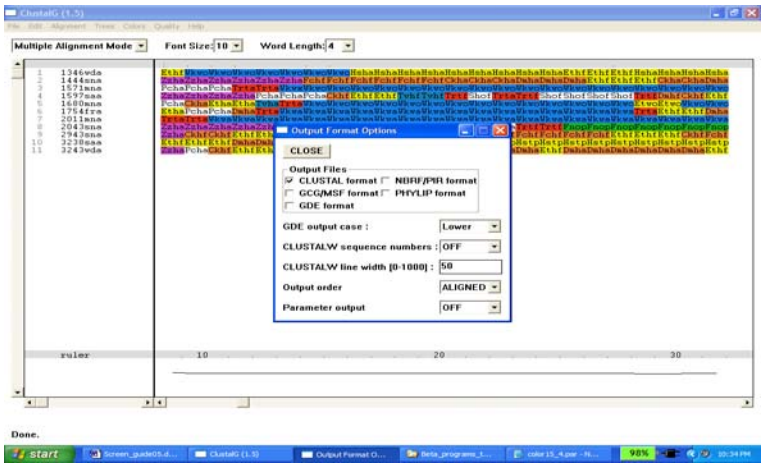
Save log file — Toggle this on with the mouse to allow the writing of the pairwise similarity values (name.lg1) and the grouping steps (name.lg2) to disk.

Log File Options — name.lg2 is a record of the joining process during the multiple alignment. The file contains a list of the sequences joined at each step. The option is to suppress the list entry unless both groups of sequences being joined contain a minimum number of sequences. If there are more than about 100 sequences, the file contains many steps that may join only a few or single sequences to a larger group. A threashold of 5 means that both groups being joined contain 5 sequences so only fairly significant joins are recorded.

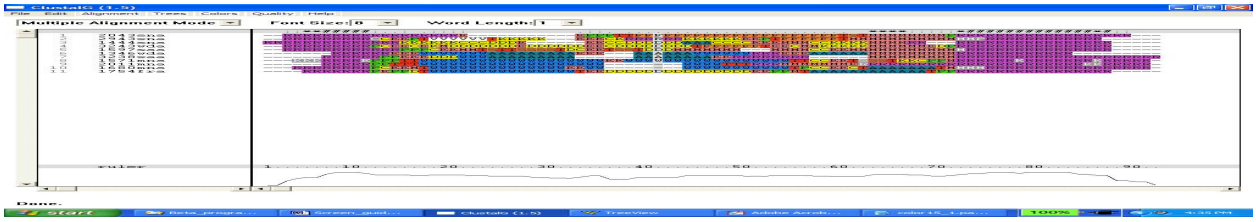
Output format options — This invokes a window that allows the user to specify the width of alignments being output, and the format to be used.



DRAFT



Output Files	User selects one or more formats for the output files. The output line length 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.
Line width	Controls the block width for printing the alignment to the ascii file name.aln. The number refers to words, not characters. Where word length is 1, these are the same.
Output Order	The aligned sequences may be written in input order or as aligned. After the alignment has been examined and saved, the input order may be preferable for final runs that generate inter-sequence distances and a new tree from the multiple alignment. The distances are then written to a matrix in their input order. This is useful if the input order had some significance such as an ordering that defined certain groups.
Parameter Output	The user may choose to have all selected parameters written to a file for future reference.

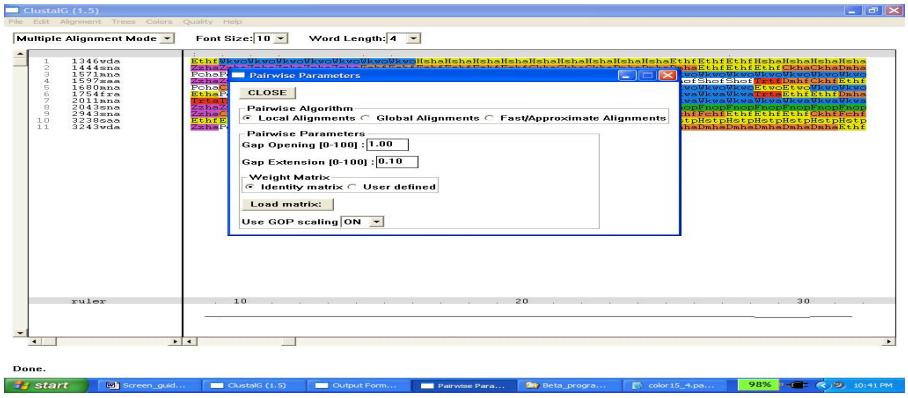


Alignment parameters sub-menus:

Reset New Gaps
Reset All Gaps — These are selected with the mouse and display a check when selected

Screen 3_a Pairwise Alignment Parameters screen

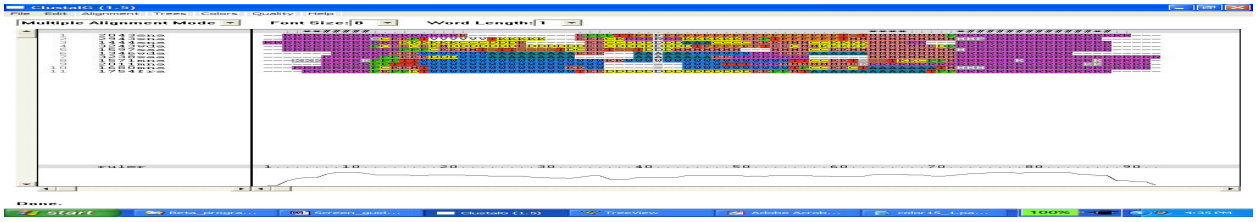
This option invokes a dialogue box that allows the user to control the set of pairwise alignments that are computed first and from which the guide tree is calculated which in turn controls the multiple alignment step. The pairwise similarities are used to generate a Neighbour-Joining Tree [8] called the guide tree that directs the order of sequence joining during the multiple alignment stage.



Pairwise algorithm — At the pairwise stage ClustalG offers local or global alignment algorithms or a faster, but less thorough string comparison process called hashing. Most personal computers are fast enough to allow use of either local or global alignment algorithms at the pairwise stage. There is some evidence that global algorithms are more powerful when the sequence similarities are low and the classification problem is difficult [9].

Pairwise parameters — The selection of gap opening and gap extension parameters determines the appearance of the final alignment and the accuracy with which pairwise similarity matrices capture the structure of sequence categories or groupings. The default values are recommended.

Weight matrix — A file containing weights for element matches and substitutions may be input. The default is to use an identity matrix with diagonal values set to

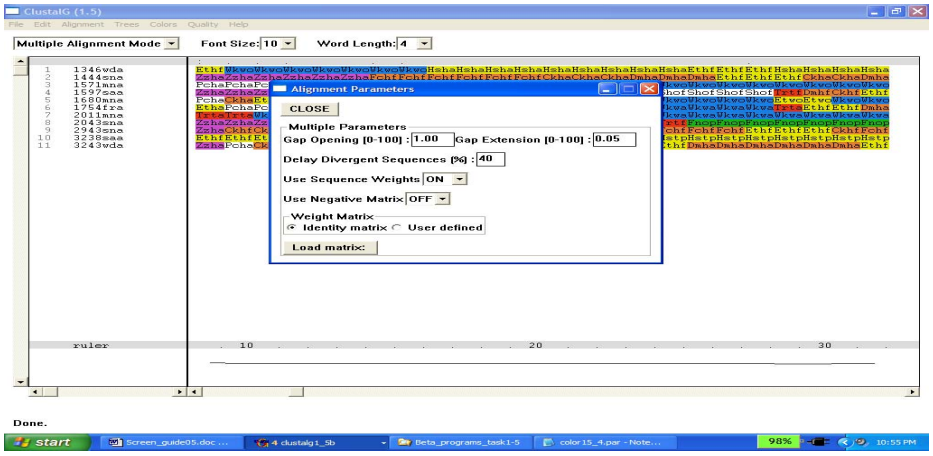


10. Gap penalty parameters should relate to weight matrix values. There is currently no strong evidence that weights other than identities are reliable in the extraction of subgroup structures. The Load Matrix button opens a file selection window. The

GOP Scaling

Gap Opening Penalties (GOP) may be scaled during alignment runs to force gaps to be grouped in roughly rectangular areas. This feature is a holdover from the biochemical applications and we recommend that it be turned off.

Screen 3_b Multiple Alignment Parameters screen



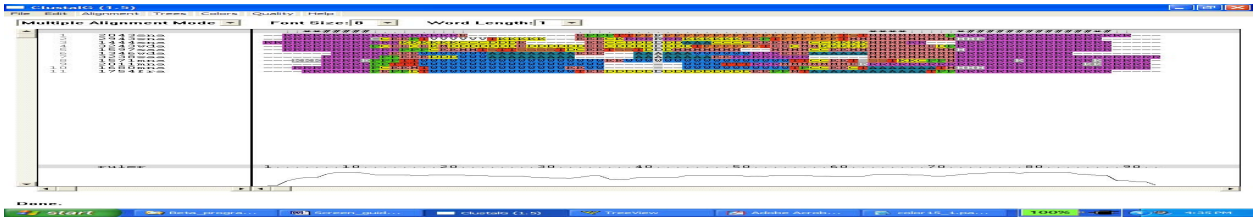
This option invokes another dialogue box that allows the user to control the set of parameters used by ClustalG in conjunction with pairwise similarity scores to calculate a guide tree, and from there to assemble the multiple alignment.

GOP GEP

Gaps may be selected with reference to the identity values of 10. Higher values will create narrower alignments that have the general appearance of the input sequences. Low values allow the alignment to spread widely removing any visual representation of the raw data. There is evidence that, at the pairwise stage, this is an advantage. However, the multiple alignments can be useful if they resemble the raw data to some extent. Suggested values are: GOP = 8 GEP = 3

Divergent Sequence

The percent identity in a pairwise alignment gives an approximate measure of similarity or divergence. A threshold can be set to exclude



any sequence that is less than the specified identity from the multiple alignment until all other eligible sequences are joined. Values of 20 to 30 percent are common in time use data, so this value should be set to about 25% or less.

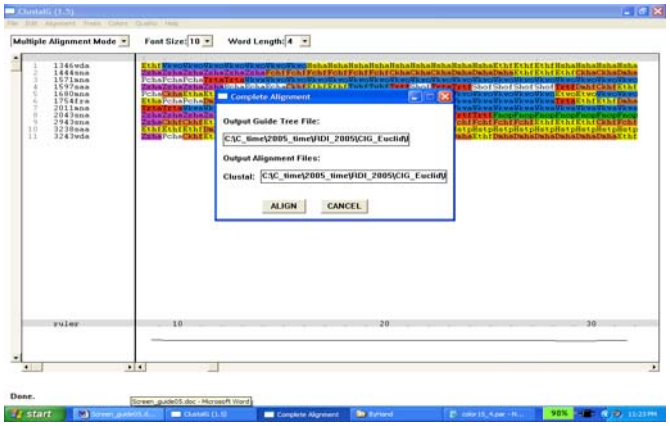
Sequence Weights — The *ClustalGByhandExample* document [10] discusses the weighting of sequences as a negative function of the branch lengths in the guide tree. Setting ON produces adequate results and there is no evidence that turning the option off is an advantage.

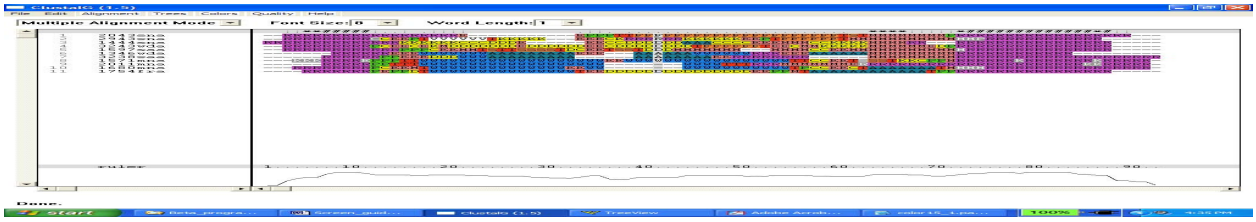
Negative Matrix — ClustalG allows users to put negative elements in the similarity matrices or parsing files. Setting the option to Off recalculates similarities to remove negative values.

Weight matrix — Usage is the same as for the pairwise screen.

Gap parameters — This opens a screen that allows specific manipulation of gap values. User Defined Penalties allows adjustment of GOP for specific activity character values. Separation Distance controls the spacing of gaps. A small value allows many narrow gaps. The default of 8 tends to produce a block appearance in the multiple alignment. End Gap Separation can be set to Off to allow unrestricted end gapping. This is appropriate when sequences have different lengths.

Screen 3_c Do Complete Alignment Screen





DRAFT

Output Guide Tree

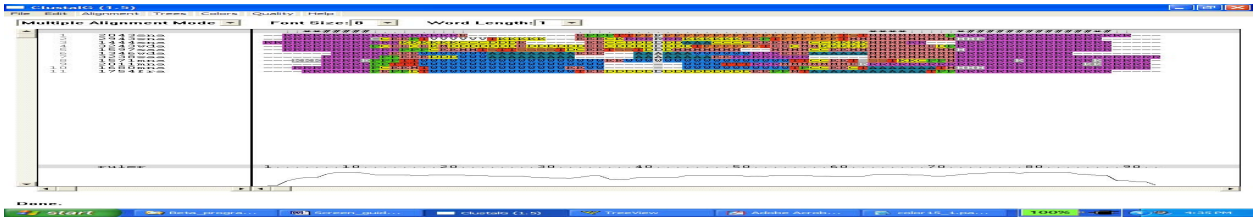
Output Alignment — This screen allows the user to name the ouput files. The default is to use the same name as the sequence file that was loaded but now using extensions of *.aln and *.dnd. An alignment file is shown later. The guide tree or dendrogram is written as a text file of nested parentheses 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 many tree drawing software packages. Treeview by Rod Page, Department of Biology, University of Glasgow is very effective. A Treeview [9] screen is shown later.

Pairwise Similarities — Writes columns of similarities, percent identities and alignment lengths to the file name.lg1, if Save Log File is toggled on.

Produce guide tree only — Writes the name.lg1, name.lg2, and name.dnd files but does not produce the multiple alignment.

Alignment from guide tree — Specifies an existing guide tree to be the basis of the multiple alignment. New multiple parameters may be selected but the pairwise selections that went into the old guide tree cannot be changed.

The following two screens show the alignment of the 11 diaries for roughly the work day, from positions 20 to 72 of the multiple alignment. Diaries of employed and domestic workers are clearly separated.



DRAFT

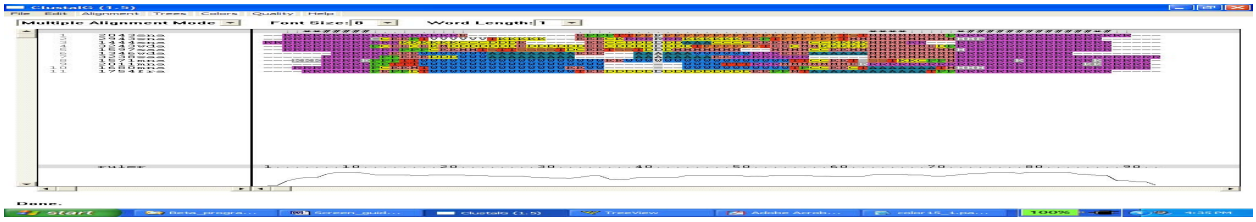
Screen 4 Trees

Output Format Option	This displays option boxes for saving tree calculation data. The tree calculations give the final distance matrix derived from the multiple alignment, and the neighbour-joining tree. The matrix file (name.phdst) will have a sequence label followed by as many rows of eight inter-sequence distances as are required by the sample size. The tree file (name.ph) will be in the nested parenthesis format. Select Phylip Tree and Phylip Matrix in the window. Check Save Log File.
Draw N-J Tree	Command opens a box to define output file names. OK launches option.

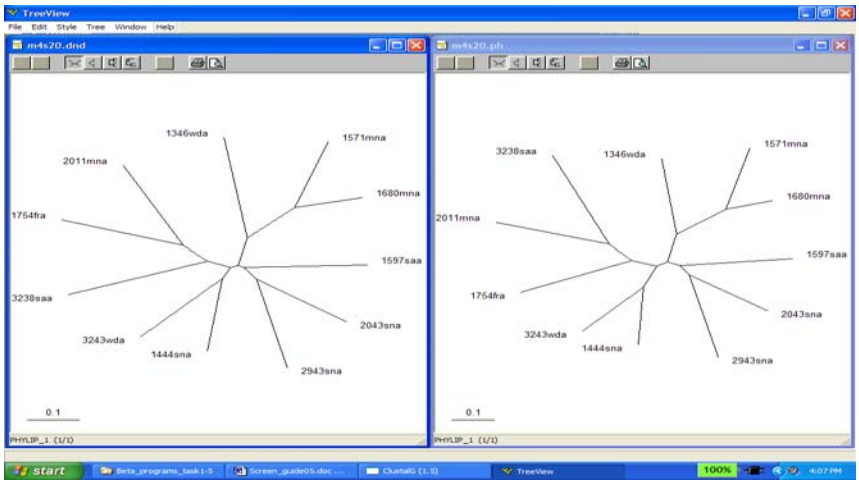
Treeview illustration of the ClustalG guide tree file

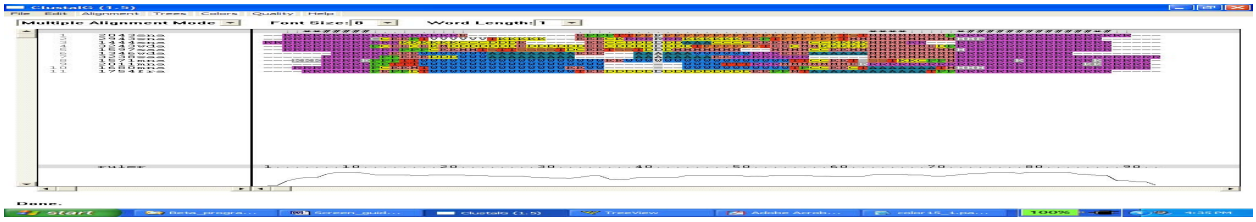
The Treeview screen illustrates the name.dnd and name.ph files. The actual tree branch lengths and node labels are written to these files. The illustrations show the calculated trees.

The trees identify three groups with identical membership. These are the first three, the second five and the last three sequences in the multiple alignment. This identity is unusual and is a consequence of the small sample size. The relative efficiency of the pairwise and multiple distance matrices has not been determined for social data. Multiple alignment distances are considered more accurate in biochemical research.



DRAFT


TreeView
m4s20.dnd
1346wda
1571mna
2011mna
1680mna
1754fra
1597saa
3238saa
2043sna
3243wda
1444sna
2943sna
0.1
PHYLIP_1 (1/1)
start
100%



DRAFT

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, *Sociological Methods and Research*, forthcoming.

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. Wilson, C. Reliability of Sequence Alignment Analysis of Social Processes: Monte Carlo Tests of ClustalG Software, *Environment and Planning A*, forthcoming

9. Saitou and Nei, Reconstructing phylogenetic trees, *Molecular Biology and Evolution*, vol. 4, no. 4, p. 406.

10. Thompson and Wilson 2005, *ClustalGByHand*, manuscript from authors.

11. Page, R. D. M. TREEVIEW: An application to display phylogenetic trees on personal computers, *Computer Applications in the Biosciences*, 12:357-358.