# Readme

Notes on ClustalG (version 1.5):
A global sequence alignment software package for application in social and natural sciences.

ClustalG is a generalized version of the Clustalx (in dos, Clustalw) originally developed by Desmond Higgins for use in the analysis of protein and nucleic acid molecules. ClustalG aligns sequences of words of up to twelve characters using the algorithms developed for previous versions of the Clustal software. ClustalG was written to offer researchers in social and natural sciences access to one of the well-known alignment software packages developed by molecular biologists, while avoiding the explicitly biochemical features of Clustalx and Clustalw. The project was undertaken as part of a research project on time use sponsored by the Social Science and Humanities Research Council of Canada.

ClustalG allows users to align words rather than letters, to construct custom similarity matrices, and to export the taxonomic groupings of input sequences which have been identified by the alignment into statistical analysis packages for multivariate analysis. These changes are additions to the basic capabilities of the biological versions of Clustal that perform global alignments, profile alignments, and calculation of taxonomic trees. ClustalG is freely available to researchers through Andy Harvey (harvey@smu.ca) of the Department of Economics, Saint Mary's University, Halifax and Clarke Wilson (wilsonjc@magma.ca).

## Element or Word Length:

The main limitation of alignment software written for molecular biology on applications in other fields is the word length. Generally, software packages use an alphabet of 20 characters (representing the 20 amino acids) plus a few characters to represent the nucleotides that comprise DNA and RNA. While some useful work has been done within this restriction, it is clear that many applications of sequence alignment or optimal matching methods in most branches of science need to represent more than 20 to 26 classes of events or conditions. ClustalG allows the user to define sequences of elements or words (each consisting of up to twelve characters) which allows the vocabulary to represent 26**12 combinations of events. The word [Wha] in a time use study could represent a work episode [W..], located at home [.h.], and which occurred when the subject was alone [..a]. ClustalG continues to allow single letter words (alphabets).

## Definition of element similarity:

The second generalization implemented in ClustalG is the removal of similarity matrices derived from amino acid substitution data and their replacement with an identity matrix as the default option. We also have an efficient means for inputting similarity scores for

comparison matrices that could be very large.

**Interface with statistical analysis software:**

The third generalization is designed to facilitate the output of sequence labels to files that may be later merged with statistical analysis systems. It appears that one useful application of sequence analysis or optimal matching will be to develop groupings or taxonomies of subjects on the basis of similar sequences of behaviour or histories. Group membership defines a nominal variable that describes a behavioural pattern derived from the analysis. Such groupings may be wanted for multivariate statistical analysis as either left hand side (dependent) or right hand side (independent) variables in regressions. As a dependent variable, group number would be employed in a logistic regression (or in an analysis of variance) that examines the influence of socio-economic characteristics in determining group membership and hence behaviour. As an independent variable, the group number would be expected to reveal the influence of the behavioural patterns of the group on some dependent variable of interest. For example, differences in patterns of homework activity may influence grades of school children independently of total time spent on homework. The nominal variable representing group membership (pattern type) appears in the regression as a dummy variable. School grade could be the dependent variable.

**ClustalG executable file:**

The zip package contains two versions of the executable. One has the standard .exe extension. The second has an extension of .eee. The second version is included as mail servers will occasionally extract and destroy .exe files.

**ClustalG Help file:**

The package of programs distributed with the ClustalG executable contains a help file. This may be read in the ClustalG window, or can be saved and printed using a word processor. The help file is the major source of documentation on ClustalG and contains references to publications describing the Clustal research program. These should be read by users interested in the details of the performance of Clustal.

**Installation:**

PC - Windows (32 bit)

The ClustalG zip file should be copied into a folder and expanded. For the testing process simply execute ClustalG1_0.exe from the Explorer window. The path to the executable must have no blanks.

**Test files:**

This version of ClustalG is distributed with one test file called c12sam.seq. This contains a six percent random sample of 790 diaries from Cycle 12 (time use) of the Statistics Canada General Social Survey, 1998. The final 141 diaries are from persons that self-identified as being disabled and contain a "D" in their respondent identifiers.

The activity codes are as follows:

| | | | |
|---|---|---|---|
| Work | W | Travel | T |
| Domestic | D | Shopping | S |
| Cooking | C | Television, radio | M |
| Personal care | P | Sleep | Z |
| Education | I | Volunteering, religious practice | V |
| Active leisure | A | Unknown | X |
| Socializing | L | | |
| | | | |

The colour parameter files color22_2.par and colprint.par are distributed with the sample files. color22_2.par is suitable for use with the sample sequences. When using the Postscript writing option, if colprint.par is missing a warning is generated but this is not usually a fatal error. Close the warning box and continue.

The Postscript facility allows the production of a pdf file and this version of ClustalG allows the file to be compressed on to one page. This is very useful for viewing large alignments.

Clarke Wilson
cwilson@cmhc-schl.gc.ca
12 January 2007

This is the last release of ClustalG. ClustalTXY is now being developed to do all of the functions of ClustalG but also to read a file of Cartesian co-ordinates and to incorporate Euclidean distance, if available, into the analysis. A beta version of this software is available.