NAME

MASE File Formats: *MATRIX, CHARACTER-SET, OUTPUT STRINGS,* and *SEQUENCE DATA FILES.*

SYNOPSIS

It is obvious that the ‘‘data’’ that *MASE* works on needs to be stored in a file. Additionally, to maintain flexibility, *MASE* reads most of it's behavioral characteristics from files external to *MASE* itself. This allows one version of the program to behave in different ways for different people, and allows individuals to easily ‘‘tune’’ *MASE* to best suit their needs. These data need have some structure; this document defines that structure.

In general, if one desires to use a structure other than those defined here, or would find an extension useful, only a single routine need be modified (if one desires to both read and write in a different format, that may well require modifying two different routines).

DESCRIPTION

SEQUENCE DATA FILES

The sequence data files use a simplified *IntelliGenetics* standard format. Compared to *GenBank* format, this is over simplified, and obscures some significant information. It is, however, simple and congenial to importing and exporting data, and to interfacing to word processors. The sequence information is maintained in it's entirety; this is, after all, our main goal.

Each data file contains one or more sequences. Each sequence contains an indefinite number of comment lines containing a semi-colon (‘‘;’’) in the first column. The first line NOT having a semi-colon in column one contains *LOCUS NAME,* a textual label for the locus. For compatibility with other software, the locus name should contain ten or fewer alphanumeric characters (most of our software will tolerate up to twenty characters in the locus name, and will tolerate blanks, tabs . . .).

Following the locus name is the sequence information. This consists of one or more lines of characters of data. White characters will be stripped from the input. There should be no more than 95 characters per line. On output, the sequence data will be formatted into lines of *SEQUENCE-LINE-LENGTH* characters each, with the white space discarded. The sequence information is terminated by the beginning of the comment block of the next sequence, or by the end of the file.

MASE provides the ability to work concurrently with sequences from several files; on saving, the sequences will be written back to the file from whence they came *(FILE-MOVE* is provided to move a locus into another file).

MASE is concerned principally with editing the sequence information. Provision is made to edit the comment lines. Currently, there is no way to change the locus name within *MASE;* this is likely to change.

See also: *LOAD* and *SAVE*

SIMILARITY MATRIX FILES

The format defined for the matrix file is weak, and will likely be rethought or at least greatly expanded. This description was valid when written.

Lines containing a semi-colon (‘‘;’’) in column 1 will be ignored, and may be used to contain comments.

Each entry line contains a definition for a set of pairs. This definition consists of three fields, separated by spaces or tabs.

The first field is a floating point number, which defines the score value for all pairwise comparisons that are to be defined.

The second field defines character set one; the third, character set two.

Pair wise comparisons for each member of set one with each member of set two is set to the score from field one.

There are two special cases. A star (‘‘*’’) is taken to mean ‘‘all characters’’. If an ampersand (‘‘&’’) composes the second field, then it is taken to mean each of the members in the first set compared to THEMSELVES. All lines in this file are interpreted sequentially; early entries may be partially overwritten by later entries.
By example:
-1.0 * *
would set ALL comparisons to have a score of ‘‘-1.0’’; this will define our mismatch score.
1.0 IVL IVL
would make IxI, IxV, IxL, VxI, VxV, VxL, LxI, LxV, and LxL all have a score of 1.0
3.0 IVL &
would make IxI, VxV, and LxL have a score of 3.0
0.0 - *
would cause a gap across from anything have a score of 0.0. Be especially cautious about the order of the assignments.

See also: *SIMILARITY*

CHARACTER DISPLAY SET FILES

MASE allows one to use an aliasing character set to display sequence information. This affects only the display; specifically, it does not effect *SAVE, OUTPUT, SEARCH, SEARCH-AGAIN* nor *PATTERN-HIGHLIGHT.*

This mode may be useful for editing alignment of protein files; one could arrange it so that ‘‘Isoleucine’’ and ‘‘Leucine’’ would appear the same on the screen.

The format of this file is extremely simple. Each line contains one ‘‘record’’. The first character of the line defines the character that will be displayed. Subsequent characters on that line are the sequence characters that should be displayed as the first one.
Thus, if one of the lines was:
iIL
then both ‘‘I’’ (Isoleucine) and ‘‘L’’ (Leucine) would be represented on the screen by an ‘‘i’’.

One may include blank lines.

Lines that have a <TAB> character in the first column will be ignored. This character was chosen since it is one of the few things that could never be displayed, since it takes more than one column position to output. This is done to permit one to include comment lines. Note that a <SPACE> character may be used - one may desire to hide certain bases all together.

See also: *CHARACTER-SET, CONSENSUS,* and *HIGHLIGHT-DIFFERENCES,* |

OUTPUT MAPPING STRINGS

The format used to define string mappings for output are somewhat awkward to use, mainly because of the complexity of the escape sequences required to generate different colors or attributes for output devices.

Lines beginning with a <TAB> character are ignored, and may be used for comment lines.

Lines containing no text are properly ignored.

The character in column one defines which character is to be mapped on output.

The remainder of the line, columns two through the end of the line, contain the string that should be printed for the letter read in column one.

This output string will be processed by *STRING CONVERSION* (see discussion in the *INTRODUCTION).*

For a special case, the string mapped to <SPACE> is used as a ‘‘reset’’ string, used to turn off all attributes (as for the ends of lines . . .)

There are two different cases of *OUTPUT*, mapping by letters and mapping by patterns. When mapping by letters occurs, the mapping string for characters encountered for output will be sent to the file instead. When mapping by patterns occurs, *OUTPUT* attempts to mimic the screen display when *PATTERN-HIGHLIGHT* has been used.

The string specified for ‘‘1’’ will be sent to enable ‘‘reverse video’’, and the string specified for ‘‘0’’ will be used to set ‘‘normal video’’.

See also: *OUTPUT* in *INTRINSIC FUNCTIONS*,
STRING CONVERSION in *INTRODUCTION*, and the discussion of *OUTPUT* in the *TUTORIAL*

EXAMPLES

FILES

SEE ALSO

BUGS