
ForCon : a software tool for the conversion of sequence alignments

Jeroen Raes and Yves Van de Peer - Department of Biochemistry, University of Antwerp (UIA), Universiteitsplein 1, B-2610 Antwerpen, Belgium*

**To whom correspondence should be addressed (E-mail: yvdp@uia.ua.ac.be)*

key words : sequence alignments – phylogenetic analysis – file conversion

Abstract

ForCon is a software tool for the conversion of nucleic acid and amino acid sequence alignments that runs on IBM-compatible computers under a Microsoft Windows environment. The program converts alignment formats used by all popular software packages for sequence alignment and phylogenetic tree inference. ForCon is available for free on request from the authors or can be downloaded via internet at URL <http://bioc-www.uia.ac.be/u/jraes/index.html>. It is also included in the software package TREECON for Windows (see <http://bioc-www.uia.ac.be/u/yvdp/index.html>).

Introduction

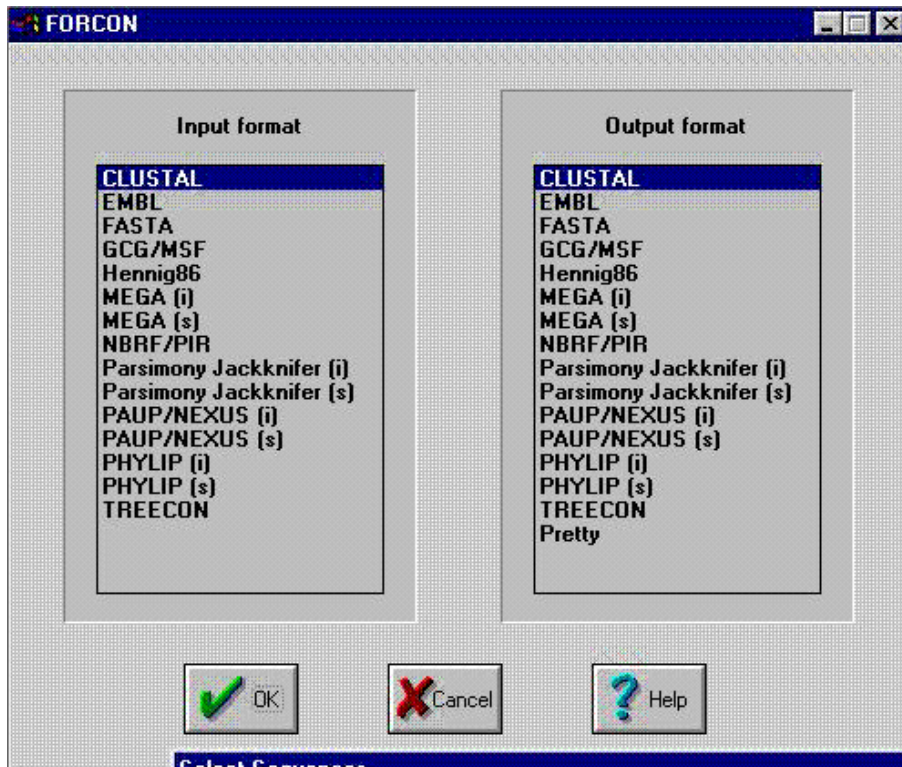
ForCon is a user-friendly software tool developed for the easy conversion of nucleic acid and amino acid sequence alignments. Sequence alignments are indispensable for many applications such as the development of probes and the inference of evolutionary trees. At the moment, many software packages for sequence alignment and the construction of evolutionary trees have implemented their own standard of saving and reading sequence alignments. Converting one alignment format into another usually requires the use of a word processor and manual

editing. To circumvent this sometimes slow and tedious work, a program was written to convert different sequence alignment formats automatically.

System requirements

ForCon is written in C++ (Borland 4.5) and runs on IBM-compatible computers operating under a Microsoft Windows environment (Win95, WinNT 3.5, 4.0 and Windows 3.1). For the latter operating system, the Windows 32 bit extension

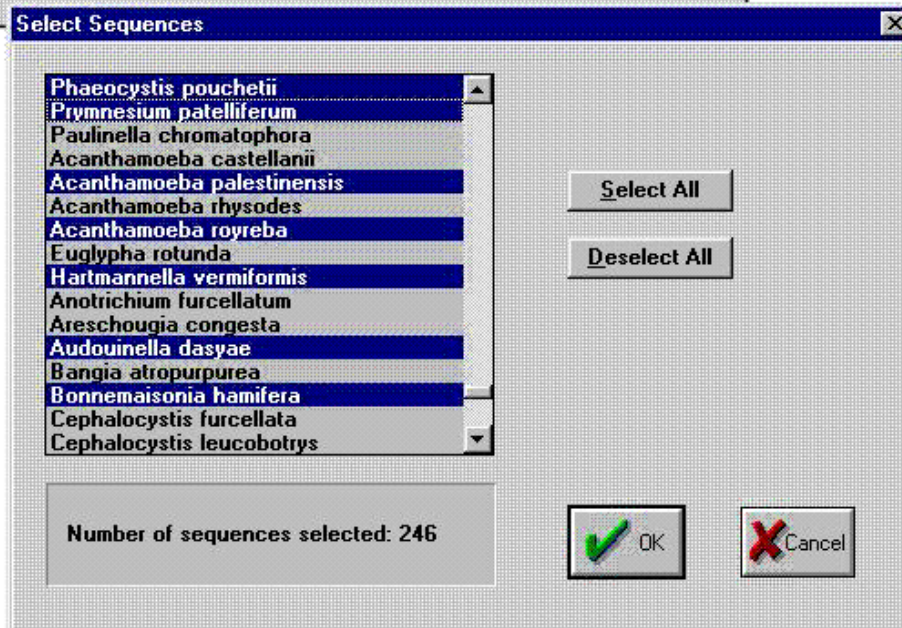
should be installed. The complete installation of ForCon requires about 1.5 Mb of hard-disk space. As dynamic memory allocation is used throughout the program, the size of the sequence alignment is constrained only by the available memory, which allows the conversion of very large files containing many hundreds of sequences. Performance tests were executed on a Pentium 200 Mhz processor with 64Mb of RAM. On this computer, an alignment of 6900 sequences of 1060 nucleotides (7.4Mb) was converted from one format into another in less than 30 seconds.



Description of the program

ForCon is developed using the Graphical User Interface provided by the MS-Windows operating system to the full extent. Although the program is very simple to use, it is assumed that users are familiar with the basic interface elements of this operating system.

At the moment, ForCon is able to convert – in both ways, i.e. reading and writing - the following formats (or formats used by the following software packages): CLUSTAL (9), EMBL (7), FASTA (6), GCG/MSF (Genetics Computer Group, Madison, USA), Hennig86 (2), MEGA (5), NBRF/PIR (1), Parsimony Jackknifer (3), PAUP/Nexus (8), PHYLIP (4), and



T R E E C O N (10). Other software packages are usually able to read one of the above mentioned formats. A so-called 'Pretty' format can also be generated for the publication of sequence alignments. When sequential and interleaved formats are supported by the original program, they are also both implemented in ForCon. Online help – with examples of the different

Upper window: alignment format selection module
Lower window: sequence name selection module

alignment formats supported – is also available.

When the program is started, it asks for the sequence alignment input file format and the required output file format (see Figure). Next, a list of all the sequences in the input file is displayed and a particular group of sequences can be selected (see Figure). This is particularly useful since additional editing of the input or output file (e.g. deleting sequences that will not be used in an analysis) can be avoided this way and different data sets can be easily made starting from the same large input file. Furthermore, it is also possible to select particular regions or codon positions from the sequence alignment and to save these only to the new file.

Discussion

To our knowledge, ForCon is the only software tool currently available that converts the large number of sequence alignment formats used by most tree construction programs. The only alternative may be the ReadSeq program developed by D.G. Gilbert (Indiana University, USA), but regarding tree construction, only the PAUP/NEXUS and PHYLIP sequence alignment formats are supported.

ForCon is available for free and can be fetched via the internet at URL <http://bioc-www.uia.ac.be/u/jraes/index.html>. The program can also be sent via electronic mail on request. Due to the structure of the program, addition of new alignment formats is very easy and users can always contact the authors if implementation of yet another format is desirable. Development of a Java™ version of ForCon has been initiated in order to guarantee platform independence in future versions of the program.

Acknowledgements

Our research is supported by the Special Research Fund of the University of Antwerp. Yves Van de Peer is Research Fellow of the Fund for Scientific Research – Flanders.

References

1. Barker, W.C., J.S. Garavelli, D.H. Haft, L.T. Hunt, C.R. Marzec, B.C. Orcutt, G.Y. Srinivasarao, L.-S.L. Yeh, R.S. Ledley, H.-W. Mewes, F. Pfeiffer and A. Tsugita. 1998. The PIR-international protein sequence database. *Nucleic Acids Res.* 26:27-32.
2. Farris, J.S. 1989. Hennig86: a PC-DOS program for phylogenetic analysis. *Cladistics* 5:163.
3. Farris, J.S., V.A. Albert, M. Källersjö, D. Lipscomb and A.G. Kluge. 1995. Parsimony Jackknifing outperforms neighbor-joining. *Cladistics* 12:99-124.
4. Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) User Manual. Department of Genetics, University of Washington, Seattle.

5. Kumar, S., K. Tamura and M. Nei. 1993. MEGA: Molecular Evolutionary Genetics Analysis User Manual. Pennsylvania State University, University Park.
6. Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85:2444-2448.
7. Stoesser, G., M.A. Moseley, J. Sleep, M. McGrowan, M. Garcia-Pastor and P. Sterk. 1998. The EMBL nucleotide sequence database. *Nucleic Acids Res.* 26:8-15.
8. Swofford, D. 1993. PAUP (Phylogenetic Analysis Using Parsimony) User Manual. Laboratory of Molecular Systematics, National Museum of Natural History, Smithsonian Institution, Washington, D.C.
9. Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin and D.G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 24:4876-4882.
10. Van de Peer, Y. and R. De Wachter. 1994. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput. Applic. Biosci.* 10:569-570.