W14–W18 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 3 May 2010*
doi:10.1093/nar/gkq321

ALTER: program-oriented conversion of DNA and protein alignments

Daniel Glez-Peña[1], Daniel Gómez-Blanco[1], Miguel Reboiro-Jato[1], Florentino Fdez-Riverola[1,*] and David Posada[2,*]

[1]Departamento de Informática, Universidad de Vigo, 32004 Ourense and [2]Departamento de Bioquímica, Genética e Inmunología, Universidad de Vigo, 36310 Vigo, Spain

Received February 3, 2010; Revised April 5, 2010; Accepted April 17, 2010

ABSTRACT

ALTER is an open web-based tool to transform between different multiple sequence alignment formats. The originality of ALTER lies in the fact that it focuses on the specifications of mainstream alignment and analysis programs rather than on the conversion among more or less specific formats. In addition, ALTER is capable of identify and remove identical sequences during the transformation process. Besides its user-friendly environment, ALTER allows access to its functionalities in a programmatic way through a Representational State Transfer web service. ALTER's front-end and its API are freely available at http://sing.ei.uvigo.es/ALTER/ and http://sing.ei.uvigo.es/ALTER/api/, respectively.

INTRODUCTION

Multiple sequence alignments (MSAs) are at the core of many bioinformatic analyses that benefit from the comparison of genomic sequences, from phylogenetic reconstruction to functional prediction (1,2). MSAs can be stored in a large variety of formats (e.g. FASTA, PIR, PHYLIP, NEXUS, etc.), and very often, researchers are obligated to transform between these in order to use different tools. Some conversion utilities have been extremely useful in this regard, the most popular being ReadSeq (http://iubio.bio.indiana.edu/soft/molbio/readseq/java/). Indeed, there are other tools developed mainly for other purposes that can also import and export aligments in several formats, like ReadAl/TrimAl (3), SeaView (4), Se-Al (http://tree.bio.ed.ac.uk/software/seal/) or even ClustalX2 (5), among others. Moreover, projects like BioPython (6) or BioPerl (7) also offer conversion capabilities.

However, the problem with most of these converters is that they—logically—focus on more or less flexible format specifications that are often violated by both developers and users. In fact, during the last years MSA's formats have 'evolved' very much like the sequences they contain, with mutational events consisting of long names, extra spaces, additional carriage returns, etc. Thus, different applications often require or produce particular MSA formats that in fact do not completely fulfill the requirements of the 'canonical' formats, often complicating the use of different tools for the analysis of data. For example, ReadSeq and programs like PAML (8) or PAUP* (http://paup.csit.fsu.edu/) fail to read simple alignments produced by ClustalX2 in PHYLIP format. To alleviate these kind of problems, we introduce a web server called ALTER for the program-oriented—rather than format-oriented—conversion between DNA and protein MSA formats. ALTER is free and open to all and there is no login requirement.

FUNCIONALITY

ALTER was designed to accomplish two main objectives: (i) easily convert between MSA formats used by popular tools and (ii) collapse sequences to haplotypes (unique sequences). In order to perform these operations in an intuitive way, ALTER implements a straightforward workflow that easily guides the user through a four-step wizard in which the different options are automatically activated when the required information is available. In addition, ALTER provides an easy-to-follow on-line help as well as many sample MSA data for testing purposes.

Program workflow

The use of ALTER typically implies four simple steps: (i) format/program identification, (ii) data load, (iii) definition

*To whom correspondence should be addressed. Tel: +34 986 812038; Fax: +34 986 812556; Email: dposada@uvigo.es
Correspondence may also be addressed to Florentino Fdez-Riverola. Tel: +34 988 387015; Fax: +34 988 387001; Email: riverola@uvigo.es

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

of conversion parameters and (iv) storage of the generated file (Figure 1).

The process of converting a given MSA in ALTER starts with the selection of the source program and/or the current format. If the user is not confident about this information, the server can try to auto detect the format of the input file.

Next, the user has to specify the operating system (OS) under which the input file was generated and upload it, or alternatively directly paste the data. In order to process the input MSA, ALTER first instantiates an appropriate sequence reader for both the input format and program. For each program/format pair, there is a specific parser generated from a formal grammar via JavaCC technology. Regardless of the possibility to reuse grammars among programs that utilize the same format, ALTER has been designed to be able to associate a different grammar for each program/format pair in order to tackle potential differences. If the user has selected the 'auto detect' option, a program-independent grammar is used instead. If there are syntax errors on the input sequences, the parser reports precise information about them and the process aborts.

Once the input MSA has been successfully read, ALTER can perform an optional step to identify redundant sequences and collapse them into haplotypes. Finally, an appropriate writer for the output program/format/OS is instantiated in order to generate the converted MSA, taking into account different parameters. These allow the user to (i) generate sequential or interleaved sequences (in NEXUS and PHYLIP formats), (ii) use lower case for residues, (iii) use match characters ('.') to indicate that the same residue is located at the same position of the first sequence and (iii) generate the sum of the number of residues at each sequence line (ALN format). In addition, the collapsing step can be configured to (i) treat gaps as missing data, (ii) consider missing data as differences between sequences and (iii) define a maximum limit of differences to collapse sequences. It is also possible to generate a program-independent conversion using only the canonical format specification.

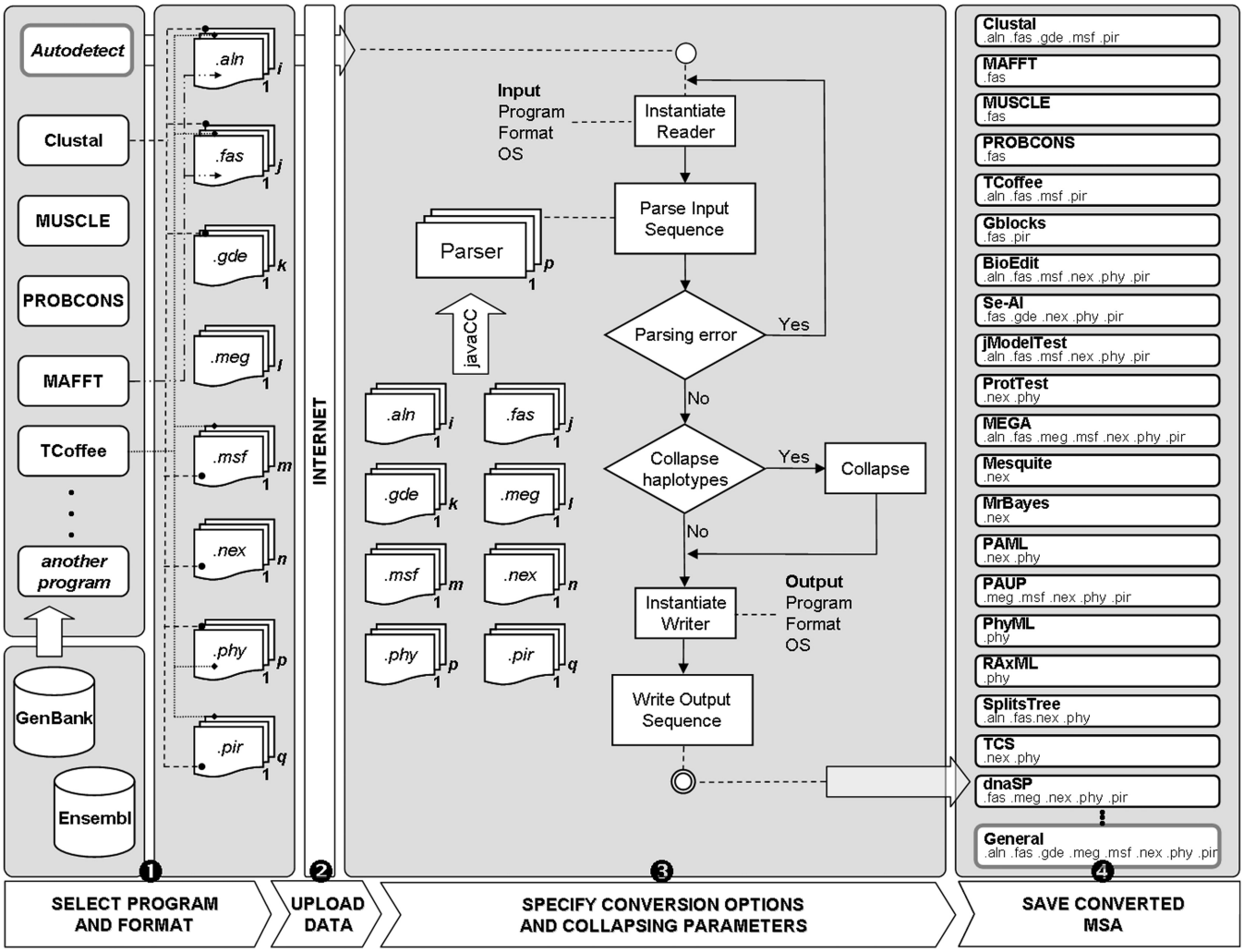

Figure 1. Schematic ALTER workflow. The user can select between different input alignment programs and formats, and obtain a MSA specifically formatted for a particular program.

Every time a new conversion job finished without errors, the output file is displayed and a download button is activated. All the relevant information related to the process of loading and recognizing the input MSA is automatically categorized (info, error, warning) and displayed to the final user by using informative log panels (Figure 2).

Supported MSA formats/programs

ALTER supports a variety of specific MSA formats provided by popular alignment tools and accepted by a variety of analysis programs. Currently, the focus is on molecular evolution, but different tools can be easily added on request. The list of programs supported include alignment, alignment filtering, sequence edition, model selection, phylogenetic, network and population genetics software (Table 1).

Web services

In addition to the functionality provided by the end user front-end, ALTER also implements a web service that allows developers to transform multiple alignment sequences directly in ALTER within their own algorithms and programs (http://sing.ei.uvigo.es/ALTER/api/). Essentially, ALTER's API offers a unique convert function with multiple parameters plus some metadata

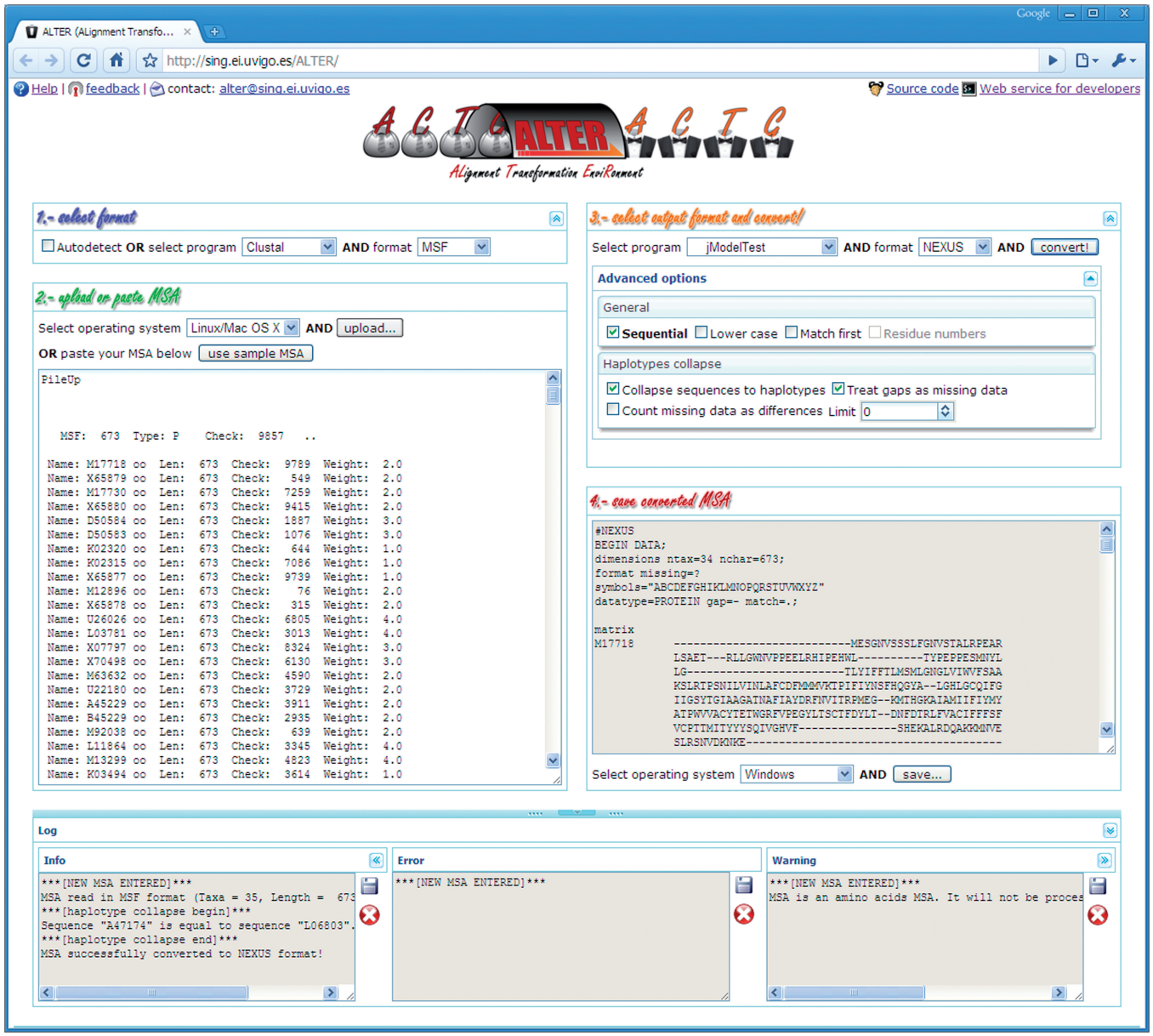

Figure 2. Example of a MSA conversion in ALTER. The 'Info panel' in the log area shows information related with the process carried out. Help, support for feedback and contact information options are available from the upper left area. Source code and a description of web services are available from the upper right area.

Table 1. List programs/formats supported by ALTER

Tools	Supported formats
INPUT: *multiple sequence alignment programs*	
Clustal (10)	ALN, FASTA, GDE, MSF, NEXUS, PHYLIP, PIR
MAFFT (11)	ALN, FASTA
TCoffee (12)	ALN, FASTA, MSF, PHYLIP, PIR
MUSCLE (13)	ALN, FASTA, MSF, PHYLIP
PROBCONS (14)	ALN, FASTA
OUTPUT: *alignment*	
Clustal	ALN, FASTA, GDE, MSF, PIR
MAFFT	FASTA
MUSCLE	FASTA
PROBCONS	FASTA
TCoffee	ALN, FASTA, MSF, PIR
OUTPUT: *alignment filtering*	
Gblocks (15)	FASTA, PIR
OUTPUT: *sequence edition*	
BioEdit (16)	ALN, FASTA, MSF, NEXUS, PHYLIP, PIR
Se-Al[a]	FASTA, GDE, NEXUS, PHYLIP, PIR
OUTPUT: *model selection*	
jModelTest (17)	ALN, FASTA, MSF, NEXUS, PHYLIP, PIR
ProtTest (18)	NEXUS, PHYLIP
OUTPUT: *phylogenetic analysis*	
MEGA (19)	ALN, FASTA, MEGA, MSF, NEXUS, PHYLIP, PIR
Mesquite[b]	NEXUS
MrBayes (20)	NEXUS
PAML (8)	NEXUS, PHYLIP
PAUP (21)	MEGA, MSF, NEXUS, PHYLIP, PIR
PhyML (22)	PHYLIP
RaxML (23)	PHYLIP
OUTPUT: *phylogenetic networks*	
SplitsTree (24)	ALN, FASTA, NEXUS, PHYLIP
TCS (25)	NEXUS, PHYLIP
OUTPUT: *population genetics*	
DnaSP (26)	FASTA, MEGA, NEXUS, PHYLIP, PIR
OUTPUT: *General*	
standard specification	ALN, FASTA, GDE, MEGA, MSF, NEXUS, PHYLIP, PIR

[a]http://tree.bio.ed.ac.uk/software/seal/.
[b]http://mesquiteproject.org/.

functions giving information about the formats and options currently supported. Table 2 summarizes the API functionality.

Supported platforms

ALTER runs on a standard Tomcat 5.5 Web application server. Currently, ALTER has been successfully tested in Internet Explorer 7, Firefox 3, Opera 9.62 and Safari 3 browsers working on Windows XP/Vista, Ubuntu Linux 8.04 version and Mac OSX 10.5 of Intel architecture.

IMPLEMENTATION

ALTER is implemented as an AJAX-enabled web application programmed in the J2SE 1.5 Java language. The ZK development framework (http://www.zkoss.org) was used to construct the user interface and to give support to JavaCC for parsing input MSA. JavaCC is a parser and a lexical analyzer generator, that is, it reads a formal description of a language (grammar) and generates code to parse instances of it. It can be see as the Java counterpart of the Lex/Flex and Yacc/Bison tools. Using JavaCC it is possible to (i) isolate the specific sequence format description in independent grammar files and (ii) generate precise error messages during parsing (9).

ALTER also implements a REST-based programming interface. Like any RESTful web service, operations are performed via web queries with a well-defined URL structure. Currently, the server gives access to the main sequence conversion functionality as well as to a set of reflective functions intended to get updated information about the supported programs and formats. This server module was implemented following the JAX-RS 1.0 (Java API for RESTful Web Services) by using the implementation found in the Apache CXF library.

Table 2. Core functionality provided by ALTER's RESTful API

Function	Description
Convert	Converts an input sequence from one format to another. This function is accessed via HTTP POST where both the sequence and parameters should be sent to the server.
Metadata functions	
List OSs	Lists the available OSs to read files from. *URL*: http://sing.ei.uvigo.es/ALTER/api/so
List input programs	Lists the currently supported input programs. *URL*: http://sing.ei.uvigo.es/ALTER/api/input/programs
List input formats	Lists the currently supported input formats. *URL*: http://sing.ei.uvigo.es/ALTER/api/input/formats
List output programs	Lists the currently supported output programs. *URL*: http://sing.ei.uvigo.es/ALTER/api/output/programs
List output formats	Lists the currently supported output formats. *URL*: http://sing.ei.uvigo.es/ALTER/api/output/formats
List output formats for a specific program	Lists the supported output formats for a given output program. *Example URL*: http://sing.ei.uvigo.es/ALTER/api/output/**paml**/formats
List options for output program and format	Lists the supported options for a given output program and format. *Example URL*: http://sing.ei.uvigo.es/ALTER/api/output/**paml**/**nexus**/options

CONCLUSIONS

Current MSA conversion tools understandably focus on the translation among 'canonical' formats, but in many instances are not of much help for users, which are interested in working with particular programs that use idiosyncratic format variations. In order to alleviate this drawback, we introduce a web server called ALTER for the program-oriented—rather than format-oriented—conversion between different DNA and protein MSA formats. In addition, ALTER is able to 'collapse' sequences to haplotypes—unique sequences—indicating which sequence corresponds to which haplotype. Eliminating this redundancy can be very helpful, for example, to speed up phylogenetic analyses.

ACKNOWLEDGEMENTS

The authors want to thank all the beta testers, especially those from the Bioinformatics and Molecular Evolution group at the University of Vigo.

FUNDING

European Research Council (ERC-2007-Stg 203161-PHYGENOM to D.P.); Spanish Ministry of Science and Education (BFU2009-08611 to D.P.); Xunta de Galicia (PGIDIT07PXIB310202PR to D.P.); INBIOMED initiative, Angeles Alvariño fellowship (to D.G-P.); University of Vigo (09VIB10 to F.F-.R.). Funding for open access charge: European Research Council (ERC-2007-Stg 203161-PHYGENOM to D.P.).

Conflict of interest statement. None declared.

REFERENCES

1. Posada,D. (ed.), (2009) *Bioinformatics for DNA sequence analysis*. Humana Press, New York, NY, USA.
2. Kemena,C. and Notredame,C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, **25**, 2455–2465.
3. Capella-Gutierrez,S., Silla-Martinez,J.M. and Gabaldon,T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
4. Gouy,M., Guindon,S. and Gascuel,O. (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, **27**, 221–224.
5. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
6. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
7. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
8. Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
9. Metsker,S.J. (2001) *Building Parsers With Java*. Addison-Wesley Professional, Boston, MA, USA.
10. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
11. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
12. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
13. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
14. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
15. Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
16. Hall,T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.*, **41**, 95–98.
17. Posada,D. (2008) jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.*, **25**, 1253–1256.
18. Abascal,F., Zardoya,R. and Posada,D. (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104–2105.
19. Tamura,K., Dudley,J., Nei,M. and Kumar,S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.*, **24**, 1596–1599.
20. Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
21. Swofford,D.L. (2000) PAUP*: Phylogenetic analysis using parsimony (*and Other Methods), Sunderland, MA, USA.
22. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
23. Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.
24. Huson,D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68–73.
25. Clement,M., Posada,D. and Crandall,K.A. (2000) TCS: a computer program to estimate gene genealogies. *Mol. Ecol.*, **9**, 1657–1659.
26. Librado,P. and Rozas,J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.