View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Crossref

Kück and Longo *Frontiers in Zoology* 2014, **11**:81
http://www.frontiersinzoology.com/content/11/1/81

FRONTIERS IN ZOOLOGY

METHODOLOGY Open Access

FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies

Patrick Kück[1*] and Gary C Longo[2]

Abstract

Background: Phylogenetic and population genetic studies often deal with multiple sequence alignments that require manipulation or processing steps such as sequence concatenation, sequence renaming, sequence translation or consensus sequence generation. In recent years phylogenetic data sets have expanded from single genes to genome wide markers comprising hundreds to thousands of loci. Processing of these large phylogenomic data sets is impracticable without using automated process pipelines. Currently no stand-alone or pipeline compatible program exists that offers a broad range of manipulation and processing steps for multiple sequence alignments in a single process run.

Results: Here we present FASconCAT-G, a system independent editor, which offers various processing options for multiple sequence alignments. The software provides a wide range of possibilities to edit and concatenate multiple nucleotide, amino acid, and structure sequence alignment files for phylogenetic and population genetic purposes. The main options include sequence renaming, file format conversion, sequence translation between nucleotide and amino acid states, consensus generation of specific sequence blocks, sequence concatenation, model selection of amino acid replacement with ProtTest, two types of RY coding as well as site exclusions and extraction of parsimony informative sites. Conveniently, most options can be invoked in combination and performed during a single process run. Additionally, FASconCAT-G prints useful information regarding alignment characteristics and editing processes such as base compositions of single in- and outfiles, sequence areas in a concatenated supermatrix, as well as paired stem and loop regions in secondary structure sequence strings.

Conclusions: FASconCAT-G is a command-line driven Perl program that delivers computationally fast and user-friendly processing of multiple sequence alignments for phylogenetic and population genetic applications and is well suited for incorporation into analysis pipelines.

Keywords: Multiple sequence alignment, Phylogenetic reconstruction, Sequence processing, Consensus sequence, Sequence translation, Sequence concatenation, File format conversion

Introduction

Phylogenetic and population genetic analyses commonly involve the manipulation and processing of multiple sequence alignments. For instance, concatenation of multiple gene alignments are common in rRNA analyses (e.g. [1-6]) and in 'mixed' nucleotide alignment analyses, combining rRNA genes like 18S and 28S as well as protein coding nucleotide sequences (e.g. [7-10]). Likewise, the ability to concatenate hundreds to thousands of nucleotide or amino acid single gene alignments has recently become an indispensable tool with the growth of phylogenomics (e.g. [11-24]). Sequence translation of nucleotide data (DNA/RNA) to protein coding sequences as well as RY coding [25] of nucleotide sequences are common practices to reduce the signal-to-noise ratio of underlying data in phylogenomic studies prior to tree reconstruction (e.g. [26-29]). In order to predict possible nucleotide sequences for a specified protein, researchers

*Correspondence: patrick_kueck@web.de
[1]Zoologisches Forschungsmuseum A. Koenig, Adenauerallee 160-163, 53113 Bonn, Germany
Full list of author information is available at the end of the article

BioMed Central

© 2014 Kück and Longo; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

often reverse translate amino acid sequences to nucleotide states (e.g. [30-33]). Another common analysis of multiple sequence alignments is consensus sequence generation, which is commonly used to identify and compare conserved and variable regions (e.g. [34-36]), design degenerated PCR primers for appropriate locations within the alignment, or to define operational taxonomic units using DNA barcode data for subsequent phylogenetic analysis (e.g. [37]). Consensus sequence generation has also become a valuable tool in large scale population genetic analyses that pool individuals as a cost effective method for determining population level data. Recent studies searching for genes potentially under selection among populations relied on identifying the most common allele at polymorphic sites as well as alleles fixed within populations [38], which can be accomplished through consensus generation.

Phylogenetic and population genetic analyses also commonly involve the tedious tasks of dealing with different sequence file formats and sequence renaming with the later becoming increasingly time-consuming when dealing with hundreds of sequences.

Although there are many scripts and online platforms that address these issues or manipulate sequence alignments with single processing steps, a software tool which enables combined processing steps in a single operation is lacking. Software like SequenceMatrix [39], TranslatorX [40], and CONCATENATOR [41] are pure concatenation tools which can be used only via graphical user interface or which are web server designed and therefore cannot be implemented in automatic process pipelines. 2matrix [42] is a pure concatenation tool as well but command line driven. SCaFoS [43] is a phylogenomic tool for selecting and concatenating sequences in large multigene and species datasets at either the amino acid or nucleotide level. Although SCaFoS is efficient at selecting orthologous sequences, creating chimerical sequences, and selecting genes according to their level of missing data, it lacks alignment processing options such as sequence translation, RY-coding, secondary structure handling, sequence renaming and consensus sequence generation.

With FASconCAT-G (FcC-G), we introduce a versatile software designed for processing and manipulating multiple sequence alignments. Conveniently, FcC-G allows for multiple processing steps in a single run and is easily implemented into pipeline analyses. FcC-G represents an advancement of FASconCAT [44], an already commonly used tool in phylogenetic studies (e.g. [45-53]).

Results and discussion

FASconCAT-G accepts multiple nucleotide, amino acid, and structure sequence alignment input files and can perform sequence renaming, file format conversion, sequence translation of nucleotide and amino acid states, consensus generation of specific sequence blocks, sequence concatenation, RY coding, model selection of amino acid replacement using ProtTest [54], extraction of parsimony informative sites as well as generation of partitioned files for MrBayes [55] and RAxML analyses [56]. The process order of FcC-G allows for a wide range of optional process combinations (Figure 1), although, some process chains are not possible in a single process run. For instance, it is not possible to RY code nucleotide sequences before translating them to amino acid sequences or to build consensus sequences before the sequence translation process. For tasks of this nature, FcC-G has to be run twice. However, we hope the current process order of FcC-G is useful for most phylogenetic and population genetic applications. To avoid errors such as the exclusion of third nucleotide site positions before sequence translation to amino acid character states, FcC-G contains a hierarchical order of single file processing steps:

1. Sequence renaming
2. Sequence translation (nucleotide to amino acid sequences or vice versa)
3. Generation of consensus sequences of predefined sequence blocks
4. RY coding of nucleotide sequences
5. Exclusion of each third nucleotide site position
6. Sequence concatenation
7. Extraction of parsimony informative sites
8. Print out of edited sequences and additional sequence information

Sequence renaming

Sequence names are often coded during the sequencing process or, if downloaded from NCBI, extended with additional information and non-alphanumeric signs which are often not allowed in downstream analysis programs. Accordingly, FcC-G can rename defined sequence names prior to file processing by using a user supplied info file, which lists, in each row, the old name delimited from the new name by a tabstop. Sequences which are not listed in the user supplied info file remain unchanged. FcC-G will print additional information of the sequence renaming process to a new outfile.

Sequence translation

FcC-G can translate standard nucleotide sequence states to amino acid characters and vice versa. For sequence translation of nucleotide data FcC-G uses the standard IUPAC triplet codes for amino acid characters. When translating amino acid states to corresponding nucleotide characters, FcC-G uses compressed IUPAC codes. Conveniently, FcC-G can recognize and handle

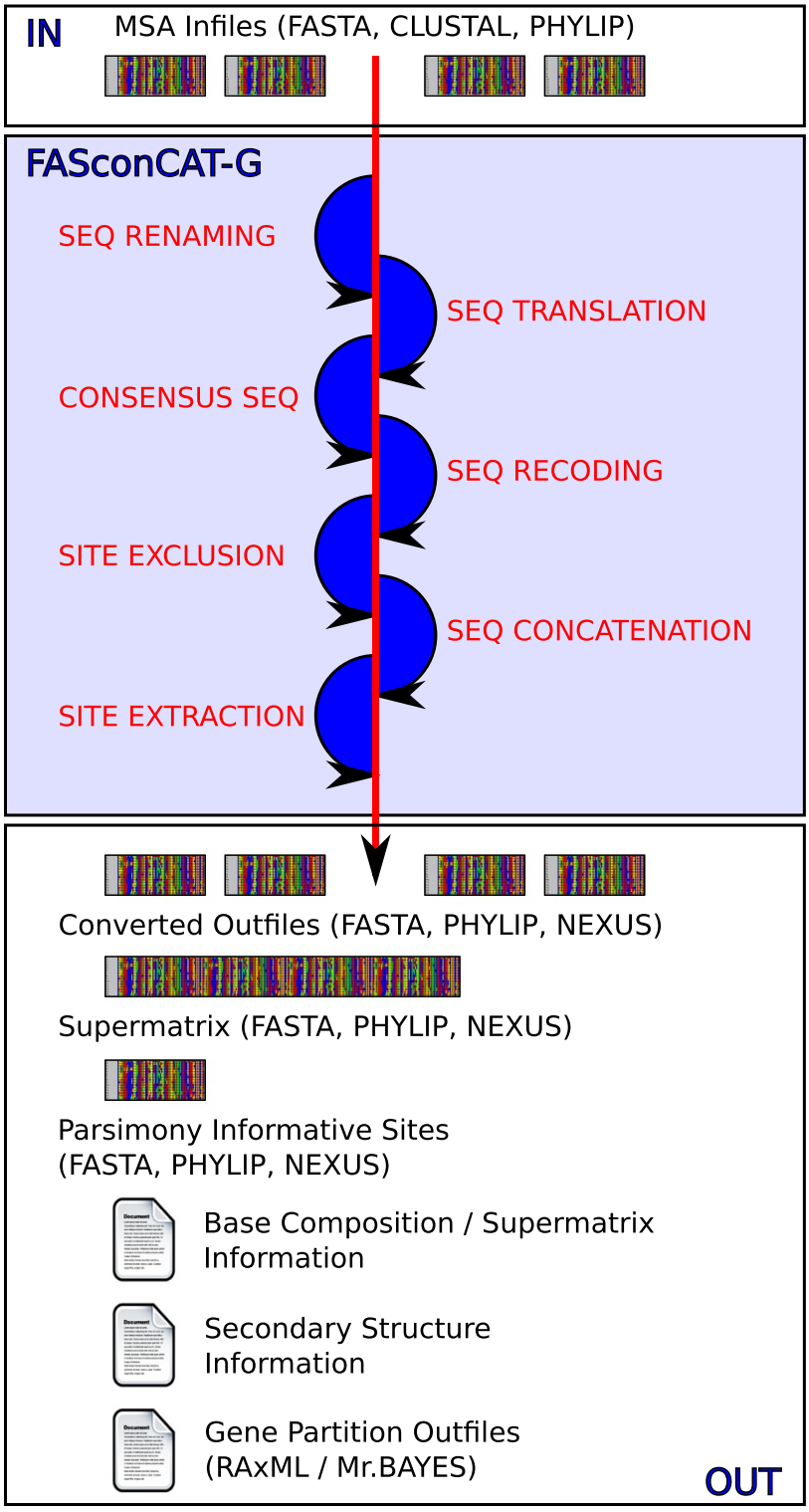

Figure 1 Simplified flowchart depicting FcC-G implemented options. Possible alignment input file formats, which can be processed individually or in combination, are listed in the top section. The hierarchical flow of possible processing options is depicted by the direction of the arrow in the middle of the figure. Processing options can be combined following the flow direction. For processing chains that contradict this flow direction, FcC-G has to be run twice. Possible output files of FcC-G are depicted on the bottom section. The content of output files depends on the chosen processing options and the type of sequences in given input file alignments. A more detailed description of possible FcC-G input/output files, implemented processing options, and examples of complex process chains are discussed in the FcC-G manual.

both amino acid and nucleotide data sets in a single processing run. Accordingly, FcC-G will only translate sequences of infiles which are suitable for a defined translation process. This makes it very easy to concatenate a mixture of different infile sequence types to one specific supermatrix sequence type FcC-G will translate incomplete nucleotide triplets to '?'. FcC-G translates nucleotide triplets even if triplets contain ambiguity codes, provided that the triplets are still assignable to specific amino acid characters (e.g. 'YTR' ↪ Leucine/L). Otherwise, unspecific triplets are translated to '?' (e.g. 'RCT' ↪ ?). FcC-G does not check for correctness of given reading frames but will print a warning in the terminal window if sequence lengths are not a multiple of three.

Consensus sequences

FcC-G can create consensus sequences for matching defined sequence blocks within given infiles using one of three consensus methods: 'Most Frequent Consensus', 'Majority Rule Consensus', and 'Strict Consensus'. The 'Most Frequent Consensus' option considers the most frequent character state at a given site among defined sequence blocks as the consensus character state. If two or more character states are equally frequent, FcC-G uses either the corresponding IUPAC ambiguity code as the consensus character state (nucleotide data) or a '?' (amino acid data and nucleotide data). The 'Majority Rule Consensus' option considers character states which occur at a given site position in more than 50% of sequences of a defined sequence block as consensus character state. Otherwise, FcC-G uses a '?' as the consensus character state (amino acid data and nucleotide data). The 'Strict Consensus' option considers all character states at a given site position to generate a strict consensus sequence for a defined sequence block using IUPAC ambiguity codes for nucleotide data and a 'X' for amino acid data. For nucleotide data, indel events (coded as '-') and missing data (coded as '?') are ignored using 'Strict Consensus' as long as a nucleotide character state exists for a specific site position. If a specific site on a defined sequence block consists of only indel events ('-') and missing data states ('?'), FcC-G will output a '?' as the consensus character state.

RY coding of nucleotide sequences

RY coding can be applied to each third nucleotide sequence position or to complete nucleotide sequences. The R code is used for purine states while the Y code is used for pyrimidines. Amino acid sequences are left unchanged unless the sequence translation option from amino acid to nucleotide states has been defined.

Sequence concatenation

FcC-G can concatenate sequence alignment infiles (nucleotide and amino acid as well as 'dot-bracket' structure information) of identical taxa into a supermatrix file. It is also possible to concatenate amino acid and nucleotide alignments into one supermatrix. In the supermatrix file, taxon sequences which were missing from single files are encoded either by 'N' (nucleotide sequences), 'X' (amino acid sequences) or by ' . ' (dots structure strings in 'dot-bracket' format).

Extraction of parsimony informative sites

FcC-G can print out additional information file(s) identifying parsimony-informative sites of given infiles and/or the concatenated supermatrix. A site is parsimony-informative if it contains at least two types of nucleotides (or amino acids), and at least two of them occur with a minimum frequency of two. The file format of parsimony-informative alignment files depends on the chosen output format(s).

Input/Output

FcC-G can simultaneously handle three different infile formats (FASTA, CLUSTAL, and PHYLIP) in any combination. Similarly, FcC-G can print concatenated and/or edited alignment files in FASTA, NEXUS, and/or PHYLIP format but FASTA is the default. NEXUS outfiles can conveniently be imbedded with MrBayes commands for direct execution in PAUP [57] or MrBayes [58] (very convenient for partitioned or mixed DNA/RNA analyses) or output without any specific commands. Likewise, PHYLIP output files can be directly used for Maximum Likelihood tree reconstruction analyses with RAxML [56] or PhyML [59]. Additionaly, our new software tool prints a file with useful information about alignment and sequence characteristics for the concatenated supermatrix as well as all single infiles. Information on this file includes single base compositions (including GC content), sequence types as well as sequence lengths and the number of taxa represented in each infile and the concatenated supermatrix. The file also contains information specific to the concatenation process, such as the position of each sequence fragment in the concatenated supermatrix as well as a list of all concatenated sequences and inserted replacement strings. However, the evaluation of this additional information often results in longer computation times depending on the size of data sets. Therefore, FcC-G offers an option to increase the overall computation speed by decreasing the information sampled and printed to the information file. If one or more infiles contain a secondary structure string, FcC-G will print another file with information about stem and loop character states and positions in both the concatenated supermatrix and infile(s). FcC-G can also print parsimony informative sites (sites which consist of at least two types of nucleotides, or amino acids, with a minimum frequency of two) from given infiles and/or the concatenated supermatrix to separate

output files. Furthermore, FcC-G can optionally generate additional gene partition output files for the concatenated supermatrix which can be directly used for Maximum Likelihood analyses using RAxML [56] or for Bayesian analyses with MrBayes [58].

Model selection of amino acid replacement using ProtTest

FcC-G offers the option to generate the best-fit protein model for each amino acid gene partition in RAxML partition formatted supermatrices using the external software, ProtTest [54]. The ProtTest option can only be

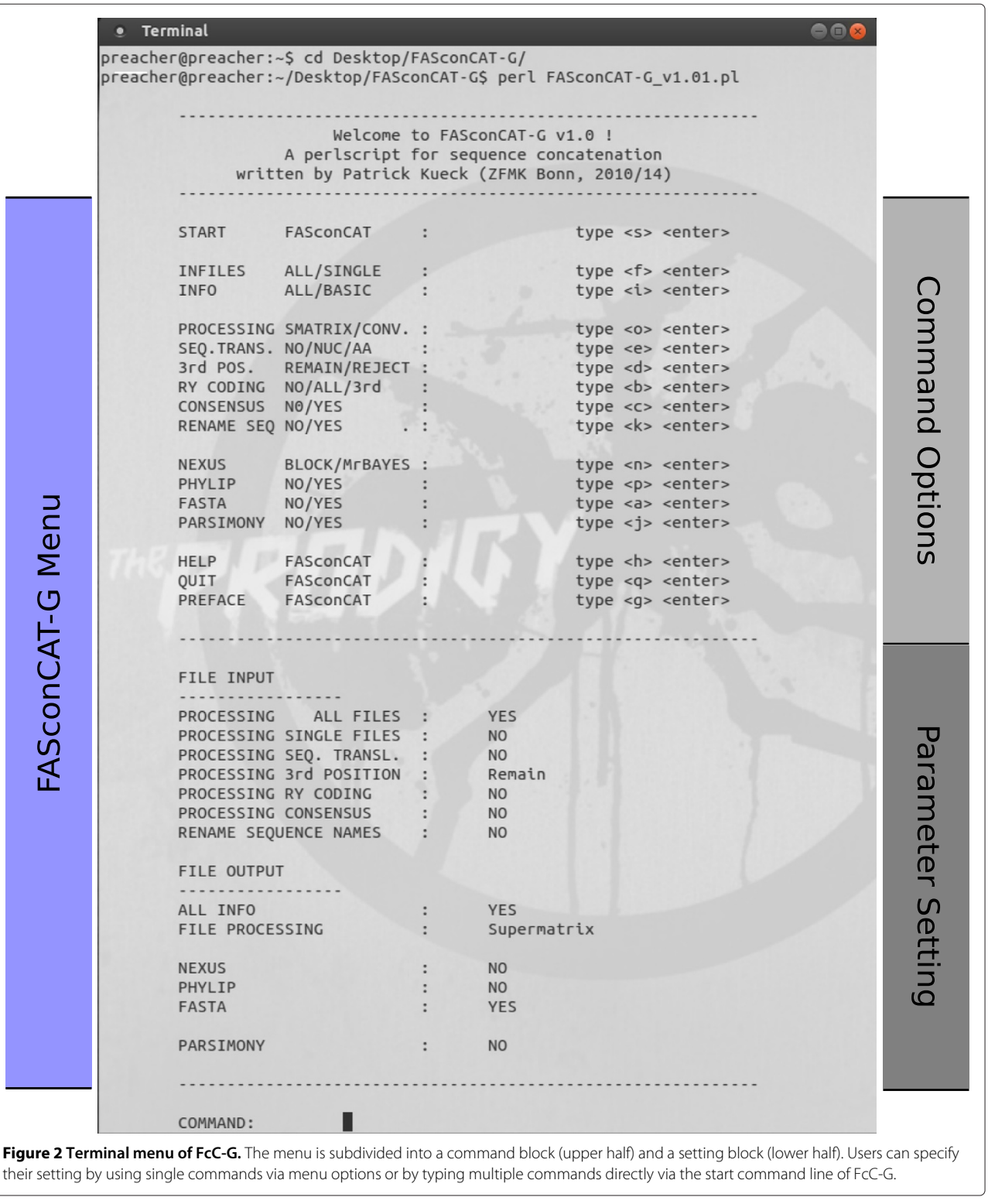

Figure 2 Terminal menu of FcC-G. The menu is subdivided into a command block (upper half) and a setting block (lower half). Users can specify their setting by using single commands via menu options or by typing multiple commands directly via the start command line of FcC-G.

executed with amino acid infiles or translated infiles and when sequence concatenation has been chosen together with the partition option ("-l"), but not for supermatrices in NEXUS format. FcC-G implements the default parameters for ProtTest version 3.3 and uses the ProtTest BIC criterium to select the best-fit model.

Conclusions

With FcC-G, we introduce an advanced editor to facilitate subsequent processing steps for multiple sequence alignments in phylogenetic and population genetic studies. Like its predecessor version, FASconCAT, FcC-G is easy to use, very fast (even with large data sets) and not limited in number of input files or input sequences. It facilitates data handling, it is time saving in generating and processing data matrices, and provides useful additional information about input sequences. FcC-G is implemented in Perl and runs on Windows PCs, Mac OS and Linux operating systems. FcC-G is command-line driven and well suited for incorporation into automatic process pipelines. Alternatively, the software tool can be operated and executed through interactive terminal menu options (Figure 2). Most processing options of FcC-G are combinable (Figure 1) and help is provided for every option. The executable source code (Additional file 1) as well as example test files and a detailed documentation of FcC-G are freely available at https://www.zfmk.de/en/research/research-centres-and-groups/fasconcat-g. The program is open-source and released under the terms of the GNU General Public License (GPL) 3.0. Detailed information and instructions are provided in the manual of FcC-G (Additional file 2). The manual also includes some practical examples, which demonstrate FcC-G is a suitable and user-friendly tool for complex phylogenetic and population genetic data processing.

Methods

FcC-G is implemented in Perl (Perl 5.0 or higher) and platform independent. Like the predecessor version FASconCAT, FcC-G can be used via command line or by terminal menu options. The terminal menu is subdivided into two parts, separated by a dashed line (Figure 2). The upper component constitutes of a list of all possible options and their associated commands for adjustment. The lower part shows the actual parameter settings of FcC-G. All default parameters can be optionally changed, and the new setting configuration will be displayed in the lower part of the menu. FcC-G is distributed under GNU GPL 3.0 and freely available from https://www.zfmk.de/en/research/research-centres-and-groups/fasconcat-g or upon request from the corresponding authors.

Additional files

Additional file 1: Executable Perl script of FASconCAT-G. FASconCAT-G is distributed under GNU GPL 3.0 and freely available. Windows users have to install a PERL interpreter on their operating system. Mac and Linux users can directly start FASconCAT-G via terminal options.

Additional file 2: FASconCAT-G manual. Detailed information and instructions and practical examples of FASconCAT-G. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PK and GCL designed FASconCAT-G. PK programmed FASconCAT-G. GCL performed the beta testing of FASconCAT-G. PK and GCL discussed and wrote the paper as well as the FASconCAT-G manual. Both authors read and approved the final manuscript.

Acknowledgements

We would like to thank all members of the Zoological Research Museum A. Koenig (Bonn, Germany) and the Center for Ocean Health (Santa Cruz, California) for inspiring discussions.

Author details

[1]Zoologisches Forschungsmuseum A. Koenig, Adenauerallee 160-163, 53113 Bonn, Germany. [2]Center for Ocean Health, 100 Shaffer Road, 95060 Santa Cruz, CA, USA.

Received: 28 July 2014 Accepted: 21 October 2014
Published online: 18 November 2014

References

1. Letsch HO, Greve C, Kück P, Fleck G, Stocsits RR, Misof B: **Simultaneous alignment and folding of 28S rRNA sequences uncovers phylogenetic signal in structure variation.** *Mol Phylogenet Evol* 2009, **53:**758–771.
2. Stocsits RR, Letsch HO, Hertel J, Misof B, Stadler PF: **Accurate and efficient reconstruction of deep phylogenies from structured RNAs.** *Nucelid Acids Res* 2009, **37:**6184–6193.
3. Keller A, Förster F, Müller T, Dandekar T, Schultz J, Wolf M: **Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees.** *Biol Direct* 2010, **5:**4.
4. Letsch HO, Kück P, Stocsits RR, Misof B: **The impact of rRNA secondary structure consideration in alignment and tree reconstruction: simulated data and a case study on the phylogeny of hexapods.** *Mol Biol Evol* 2010, **27**(11):2507–2521.
5. Murienne J, Edgecombe G, Giribet G: **Including secondary structure, fossils and molecular dating in the centipede tree of life.** *Mol Phylogenet Evol* 2010, **57:**301–313.
6. Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY: **Understanding the transcriptome through RNA structure.** *Nat Rev Genet* 2011, **12:**641–655.
7. Dinapoli A, Klussmann-Kolb A: **The long way to diversity – Phylogeny and evolution of the Heterobranchia (Mollusca:Gastropoda).** *Mol Phyogenet Evol* 2010, **55:**60–76.
8. Hoppenrath M, Leander BS: **Dinoflagellate phylogeny as inferred from heat shock protein 90 and ribosomal gene sequences.** *PLoS ONE* 2010, **5**(10):e13220.
9. Goto R, Okamoto T, Ishikawa H, Hamamura Y, Kato M: **Molecular phylogeny of echiuran worms (phylum: annelida) reveals evolutionary pattern of feeding mode and sexual dimorphism.** *Plos ONE* 2013, **8**(2):e56809.
10. Lopez-Osorio F, Pickett KM, Carpenter JM, Ballif BA, Agnarsson I: **Phylogenetic relationships of yellowjackets inferred from nine loci (Hymenoptera: Vespidae, Vespinae, *Vespula and Dolichovespula*).** *Mol Phylogenet Evol* 2014, **73:**190–201.
11. Phillipe H, Lartillot N, Brinkmann H: **Multigene Analyses of Bilaterian Animals Corroborate the Monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia.** *Mol Biol Evol* 2005, **22**(5):1246–1253.

12. Savard J, Tautz D, Richards S, Weinstock GM, Gibbs RA, Werren JH, Tettelin H, Lercher MJ: **Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects.** *Genome Res* 2006, **16:**1334–1338.
13. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SH, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler W, Martindale MQ, Giribet G: **Broad phylogenomic sampling improves resolution of the animal tree of life.** *Nature* 2008, **452:**745–749.
14. Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguna J, Bailly X, Jondellus U, Wiens M, Müller WEG, Seaver E, Wheeler WC, Martindale MQ, Giribet G, Dunn CW: **Assessing the root of bilaterian animals with scalable phylogenomic methods.** *Proc R Soc B* 2009, **276**(1677):4261–4270.
15. Simon S, Strauss S, von Haeseler A, Hadrys H: **A phylogenomic approach to resolve the basal pterygote divergence.** *Mol Biol Evol* 2009, **26**(12):2719–2730.
16. Meusemann K, von Reumont BM, Simon S, Roeding F, Kück P, Strauss S, Ebersberger I, Walzl M, Pass G, Breuers S, Achter V, von Haeseler A, Burmester T, Hadrys H, Wägele JW, Misof B: **A phylogenomic approach to resolve the arthropod tree of life.** *Mol Biol Evol* 2010, **27**(11):2451–2464.
17. Pick KS, Phillipe H, Schreiber F, Erpenbeck D, Jackson DJ, Wrede P, Wiens M, Alie A, Morgenstern B, Manuel M, Wörheide G: **Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships.** *Mol Biol Evol* 2010, **27**(9):1983–1987.
18. Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ: **A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata.** *Proc R Soc B* 2010, **278:**298–306.
19. Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B, Halanych KM: **Phylogenomics reveals deep molluscan relationships.** *Nature* 2011, **477:**452–456.
20. Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SCS, Rouse GW, Giribet G, Dunn CW: **Resolving the evolutionary relationships of molluscs with phylogenomic tools.** *Nature* 2011, **480:**364–367.
21. Struck TH, Paul C, Hill N, Hartmann S, Hösel C, Kube M, Lieb B, Meyer A, Tiedemann R, Purschke G, Bleidorn C: **Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of spiralia.** *Nature* 2011, **471:**95–98.
22. Rubin BER, Gee RH, Moreau CS: **Inferring phylogenies from RAD sequence data.** *PLoS ONE* 2012, **7**(4):e33394.
23. Wagner CE, Keller I, Wittwer S, Selz OM, Mwalko S, Greuter L, Sivasundar A, Seehausen O: **Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation.** *Mol Ecol* 2013, **22**(3):787–798.
24. Wheat CW, Wahlberg N: **Phylogenomic insights into the cambrian explosion, the colonization of land and the evolution of flight in arthropoda.** *Syst Biol* 2013, **62:**93–109.
25. Woese CR, Achenbach L, Rouvier P, Mandelco L: **Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus Fulgidus in light of certain composition-induced artifact.** *Syst Appl Microbiol* 1991, **14:**364–371.
26. Phillips MJ, Delsuc F, Penny D: **Genome-scale phylogeny and the detection of systematic biases.** *Mol Biol Evol* 2004, **21**(7):1455–1458.
27. Harshman J, Braun EL, Braun MJ, Huddleston CJ, Bowie RCK, Chojnowski JL, Hackett SL, Han KL, Kimball RT, Marks BD, Miglia KJ, Moore WS, Reddy S, Sheldon FH, Steadman DW, Steppan SJ, Witt CC, Yuri T: **Phylogenomic evidence for multiple losses of flight in ratite birds.** *Proc Natl Acad Sci U S A* 2008, **105**(36):13462–13467.
28. White NE, Phillips MJ, Gilbert TP, Alfaro-Nunez A, Willerslev E, Mawson PR, Spencer PBS, Bunce M: **The evolutionary history of cockatoos (Aves: Psittaciformes: Cacatuidae).** *Mol Phylogenet Evol* 2011, **59**(3):615–622.
29. Chen JN, Lopez A, Lavoue S, Miya M, Chen WJ: **Phylogeny of the Elopomorpha (Teleostei): Evidence from six nuclear and mitochondrial markers.** *Mol Phylogenet Evol* 2014, **70:**152–161.
30. Burger TD, Shao R, Beati L, Miller H, Barker SC: **Phylogenetic analysis of ticks (Acari: Ixodida) using mitochondrial genomes and nuclear rRNA genes indicates that the genus Amblyomma is polyphyletic.** *Mol Phylogenet Evol* 2012, **64:**45–55.
31. Liu GH, Wu CH, Song HQ, Wei SJ, Xu MJ, Lin RQ, Zhao GH, Huang SY, Zhu XQ: **Insect phylogenomics: results, problems and the impact of matrix composition.** *Mol Phylogenet Evol* 2012, **492:**110–116.
32. Lin RQ, Qiu LL, Liu GH, Wu XY, Weng YB, Xie WQ, Hou J, Pan H, Yuan ZG, Zou FC, Hu M, Zhu XQ: **Characterization of the complete mitochondrial genomes of five Eimeria species from domestic chickens.** *Mol Phylogenet Evol* 2012, **480:**28–33.
33. dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ, Yang Z: **Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny.** *Proc R Soc B* 2012, **279:**3491–3500.
34. Krüger M, Krüger C, Walker C, Stockinger H, Schüssler A: **Phylogenetic reference data for systematics and phylotaxonomy of arbuscular mycorrhizal fungi from phylum to species level.** *New Phytologist* 2012, **139:**970–984.
35. Waheed Y, Saeed U, Anjum S, Afzal MS, Ashraf M: **Development of global consensus sequence and analysis of highly conserved domains of the HCV NS5B protein.** *Hepat Mon* 2012, **12**(9):e6142.
36. Cotton M, Lam TT, Watson SJ, Palser AL, Petrova V, Grant P, Pybus OG, Rambaut A, Guan Y, Pillay D, Kellam P, Nastouli E: **Full-Genome Deep Sequencing and Phylogenetic Analysis of Novel Human Betacoronavirus.** *Emerg Infect Dis* 2013, **19**(5):736–742.
37. Blaxter M, Mann J, Chapman F, Thomas F, Whitton RF, Abebe E: **Defining operational taxonomic units using DNA barcode data.** *Phil Trans R Soc B* 2005, **360**(1462):1935–1943.
38. Rubin CJ, Zody MC, Eriksson J, Meadows RS, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S, Hallböök F, Besnier F, Carlborg O, Bed'hom B, Tixier-Boichard M, Jensen P, Siegel P, Lindblad-Toh K, Andersson L: **Whole-genome resequencing reveals loci under selection during chicken domestication.** *Nature* 2010, **464:**587–591.
39. Vaidya G, Meier R: **SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information.** *Cladistics* 2011, **27:**171–180.
40. Abascal F, Zardoya R, Telford MJ: **TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations.** *Nucl Acids Res* 2010, **38**(Web Server issue):W7—13. doi:10.1093/nar/gkq291.
41. Pina-Martins F, Paulo OS: **CONCATENATOR: sequence data matrices handling made easy.** *Mol Ecol Resour* 2008, **8:**1254–1255.
42. Salinas NR, Little DP: **2MATRIX: A utility for indel coding and phylogenetic matrix concatenation.** *Appl Plant Sci* 2014, **2:**1300083.
43. Roure B, Rodriguez-Ezpeleta N, Phillipe H: **SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics.** *BMC Evol Biol* 2007, **7:**S2.
44. Kück P, Meusemann K: **FASconCAT: Convenient handling of data matrices.** *Mol Phylogenet Evol* 2010, **56:**1115–1118.
45. Kück P, Hita-Garcia F, Misof B, Meusemann K: **Improved phylogenetic analyses corroborate a plausible position of Martialis Heureka in the ant tree of life.** *PLoS ONE* 2011, **6**(6):e21031.
46. Biswal KD, Debnath M, Kumar S, Tandon P: **Phylogenetic reconstruction in the Order Nymphaeales: ITS2 secondary structure analysis and *in silico* testing of maturase k *(matK)* as a potential marker for DNA bar coding.** *BMC Bioinformatics* 2012, **13:**16.
47. Boumans L, Baumann RW: **Amphinemura palmeni is a valid Holarctic stonefly species (Plecoptera: Nemouridae).** *Zootaxa* 2012, **3537:**59–75.
48. Kohn AB, Citarella MR, Kocot KM, Bobkova YV, Halanych KM, Moroz LL: **Rapid evolution of the compact and unusual mitochondrial genome in the ctenophore, Pleurobrachia bachei.** *Mol Phylogenet Evol* 2012, **63:**203–207.
49. McNulty SN, Mullin AS, Vaughan JA, Tkach VV, Weil GJ, Fischer PU: **Comparing the mitochondrial genomes of Wolbachia-dependent and independent filarial nematode species.** *BMC Genomics* 2012, **13:**145.
50. Young ND, Jex AR, Li B, Liu S, Yang L, Xiong Z, Li Y, Cantacessi C, Hall RS, Xu X, Chen F, Wu X, Zerlotini A, Oliveira G, Hofmann A, Zhang G, Fang X, Kang Y, Campbell BE, Loukas A, Ranganathan S, Rollinson D, Rinaldi G, Brindley PJ, Yang H, Wang J, Wang J, Gasser RB: **Whole-genome sequence of Schistosoma haematobium.** *Nat Genet* 2012, **44:**221–225.
51. Golombek A, Tobergte S, Nesnidal P, Purschke G, Struck T: **Mitochondrial genomes to the rescue – Diurodrilidae in the mystozomid trap.** *Mol Phylogenet Evol* 2013, **68**(2):312–326.
52. Larriba E, Jaime MDLA, Carbonell-Caballero J, Conesa A, Dopazo J, Nislow C, Martin-Nieto J, Lopez-Llorca LV: **Sequencing and functional analysis**

of the genome of a nematode egg-parasitic fungus, Pochonia chlamydosporia. *Fungal Genet Biol* 2014, **65:**69–80.
53. Scheel BM, Hausdorf B: **Dynamic evolution of mitochondrial ribosomal proteins in Holozoa.** *Mol Phylogenet Evol* 2014, **76:**67–74.
54. Darriba D, Guillermo LT, Doallo R, Posada D: **ProtTest 3: fast selection of best-fit models of protein evolution.** *Bioinformatics* 2011, **27:**1164–1165.
55. Ronquist F, Huelsenbeck J: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19**(12):1572–1574.
56. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**(21):2688–2690.
57. Swofford D: *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0.* Sunderland, MA: Sinauer Associates; 2003.
58. Huelsenbeck J, Ronquist F: **MrBayes: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**(8):754–755.
59. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **PhyML 3.0: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010, **59**(3):307–321.

doi:10.1186/s12983-014-0081-x
Cite this article as: Kück and Longo: FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology* 2014 **11**:81.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central