Mycoscience (2011) 52:278–282
DOI 10.1007/s10267-010-0086-z

NOTE

A note on the incidence of reverse complementary fungal ITS sequences in the public sequence databases and a software tool for their detection and reorientation

R. Henrik Nilsson · Vilmar Veldre · Zheng Wang · Martin Eckart · Sara Branco · Martin Hartmann · Christopher Quince · Anna Godhe · Yann Bertrand · Johan F. Alfredsson · Karl-Henrik Larsson · Urmas Kõljalg · Kessy Abarenkov

Received: 29 June 2010 / Accepted: 20 October 2010 / Published online: 23 November 2010

Abstract Reverse complementary DNA sequences—sequences that are inadvertently cast backward and in which all purines and pyrimidines are transposed—are not uncommon in sequence databases, where they may introduce noise into sequence-based research. We show that about 1% of the public fungal ITS sequences, the most commonly sequenced genetic marker in mycology, are reverse complementary, and we introduce an open source software solution to automate their detection and reorientation. The MacOSX/Linux/UNIX software operates on public or private datasets of any size, although some 50 base pairs of the 5.8S gene of the ITS region are needed for the analysis.

Keywords DNA barcoding · Environmental sampling · Hidden Markov models · Quality assessment · Sequence identification

The inconspicuous nature of fungal life is a challenge to the pursuit of mycological knowledge. It is now known that the fungal fruiting bodies and other propagation structures found at any collection site typically represent only a biased fraction of the full fungal diversity at that locality (Porter et al. 2008; Suzuki and Bärlocher 2009). Similarly, many data have emerged to show that morphology is an imprecise source of information for species delimitation and phylogenetic inference in most groups of fungi (Taylor et al. 2000; Hibbett 2007). Molecular (DNA sequence)

Electronic supplementary material The online version of this article (doi:10.1007/s10267-010-0086-z) contains supplementary material, which is available to authorized users.

R. H. Nilsson (✉) · V. Veldre · U. Kõljalg · K. Abarenkov
Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai St, 51005 Tartu, Estonia
e-mail: henrik.nilsson@dpes.gu.se

R. H. Nilsson · Y. Bertrand
Department of Plant and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden

Z. Wang
Department of Ecology and Evolutionary Biology, Yale University, POB 208106, 165 Prospect Street, New Haven, CT 06520-8106, USA

M. Eckart
Jena Microbial Resource Collection, Department of Molecular and Applied Microbiology, HKI, University of Jena, Neugasse 25, 07743 Jena, Germany

S. Branco
Forschungsinstitut Senckenberg, BiK-F, Senckenberganlage 25, 60325 Frankfurt, Germany

M. Hartmann
Department of Microbiology and Immunology, Life Sciences Centre, University of British Columbia, 4504-2350 Health Sciences Mall, Vancouver, BC V6T 1Z3, Canada

C. Quince
Department of Civil Engineering, Glasgow University, Glasgow G128LT, United Kingdom

A. Godhe
Department of Marine Ecology, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden

J. F. Alfredsson
Oepir Consulting, Vasagatan 48:1, 411 37 Göteborg, Sweden

K.-H. Larsson
The Mycological Herbarium, Natural History Museum, University of Oslo, P.O. Box 1172, Blindern, 0318 Oslo, Norway

information was therefore readily adopted by the mycological community as a high-resolution source of data for taxonomic and ecological research, and the past 10 years have witnessed the standardization of sequence data as a research implement throughout mycology (Blackwell et al. 2006). The most commonly sequenced genetic marker for addressing research questions at and below the fungal genus level is the internal transcribed spacer (ITS) region of the nuclear ribosomal repeat unit (Seifert 2009; Begerow et al. 2010; Eckart et al. 2010).

These advantages notwithstanding, sequence data do not always form a straightforward component of mycological research. Fungal ITS sequences with incorrect taxonomic annotation accumulate in the public sequence databases, and entries that are chimeric or that feature poorly read regions are similarly concerns (Bidartondo et al. 2008; Ryberg et al. 2009; Nilsson et al. 2010). Another problematic aspect of publicly available fungal ITS sequences is that a portion of them are incorrectly deposited in the reverse complementary orientation (i.e., backward and with all purines and pyrimidines transposed, e.g., CTAGG instead of the correct CCTAG). This error typically happens during the sequence assembly step when the sequencing software or researcher fails to relate the orientation of the sequences under assembly to that of the others being generated. Although it is trivial to reorient such sequences using any of a number of web resources (e.g., http://www.bioinformatics.org/sms/rev_comp.html) or BioPerl (Stajich et al. 2002), the process of detecting them in the first place is not always equally straightforward. And, although many software packages account for the presence of reverse complementary entries (e.g., the sequence similarity suite BLAST; Altschul et al. 1997), others are less well positioned to do so (such as programs for multiple alignment, phylogenetic inference, and—in part—sequence clustering). In manual research efforts, where reverse complementary sequences tend to stand out enough to be easy to identify, their presence often forms little more than an annoying obstacle whose resolution causes unnecessary delay. Given the increasingly integrated and automated nature of sequence-based research (cf. Hibbett et al. 2005; Kauff et al. 2007; Hartmann et al. 2010), however, human intervention preventing such entries from entering automated analysis pipelines can no longer be expected to always occur.

The present study introduces a command-line MacOS X/Linux/UNIX open source software package (Supplementary Item 1; http://www.emerencia.org/reversecomplementary.html) for detection and reorientation of reverse complementary fungal ITS sequences from arbitrarily large public or private datasets in the FASTA format (Pearson and Lipman 1988). The software is written in Perl and uses hidden Markov models (HMMs) computed from the very conserved 5.8S gene of the ITS region to examine the orientation of the query sequences. The software tries to detect a region corresponding to the first 54, very conserved, base pairs (bp) of the 5.8S (5′-end) using HMMER ver. 2.3.2 (Eddy 1998) and the HMMs of Nilsson et al. (2008); if this is unsuccessful, it attempts to locate the same region in the reverse complement of the query sequence. If the region was found in the default orientation of the query sequence, the query as given by the user is in the standard 5′–3′ orientation; if the region was found in the reverse complement of the query sequence, the query is reverse complementary. If the region could not be found in either of the attempts, the program lacks the data needed to decide upon the orientation of the query sequence.

The outdata consists of a set of files designed to let the users move on with their research swiftly. One file lists the names of the query sequences found to be given in the correct orientation; one lists the names of the queries found to be given in the reverse complementary orientation; and one lists the sequences for which the 5.8S could not be found such that a decision on the orientation of the sequence was not possible. A separate file in the FASTA format contains *all* query sequences provided as indata to the program; in this file, the entries found to be reverse complementary are given in the correct orientation. Optionally, the user can let the software compare each query sequence to a local (bundled) copy of all ~83,000 reasonably full-length, fully identified fungal ITS sequences in the International Nucleotide Sequence Databases (INSD; Sayers et al. 2010) using BLAST. The output would then also comprise a list of the most similar INSD sequences and the associated match statistics for each query sequence. In addition, the 20 (default) best matches are aligned to the query (and, for the entries found to be reverse complementary, the respective reverse complement) using MAFFT (Katoh and Toh 2008) to provide the user with a visualization of the results of the analyses (and reorientation as applicable; Fig. 1).

To estimate the proportion of reverse complementary fungal ITS sequences in INSD and to evaluate the performance of the software, all 161,922 such sequences in INSD as compiled by *emerencia* (Nilsson et al. 2005) in April 2010 were provided as input to the program. A full 149,483 (92.3%) were reported by the program to be in the correct orientation; another 1,443 (0.9%) were reverse complementary; and for 10,996 sequences (6.8%), the 5.8S could not be located, excluding these from any decision on sequence orientation. The multiple alignments of all 1,443 reportedly reverse complementary entries, as well as the alignments of 1,000 random entries each from the two other categories, were scrutinized manually (Supplementary Item 2). Upon inspection, all 1,000 alignments (100%) of queries, indicated as being in the correct orientation,

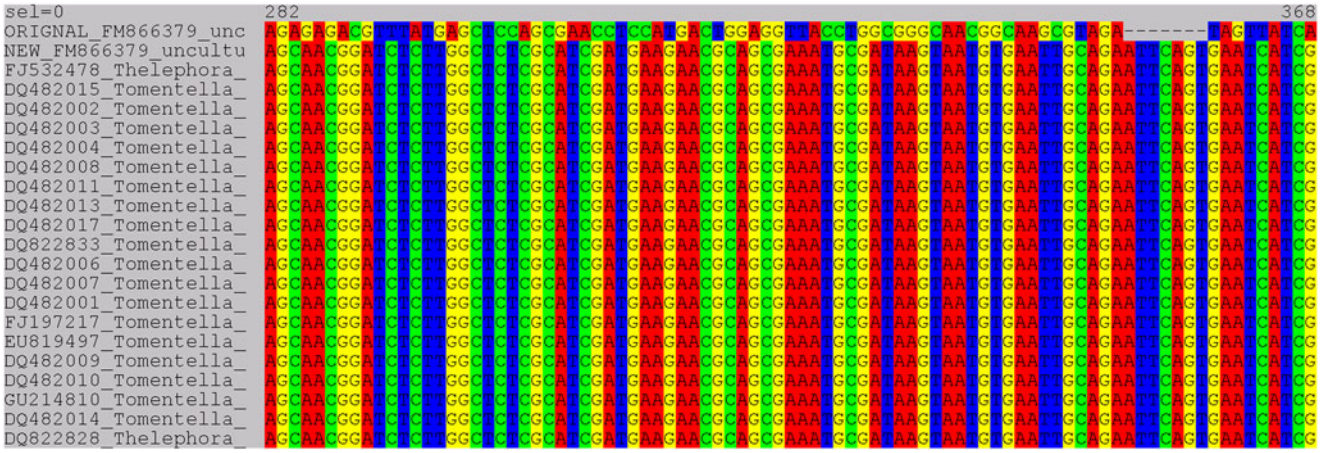

Fig. 1 Multiple alignment of a reverse complementary query sequence as deposited in INSD (*topmost sequence*) and as reoriented by the present software (*second sequence from top*). The remaining 20 sequences are the closest BLAST matches of the query and are given in the order listed by BLAST. SeaView (Gouy et al. 2010) was used to display the alignment

showed the query to be correctly oriented as deposited in INSD. Similarly, all 1,443 alignments (100%) of queries indicated as being in the reverse complementary orientation showed that the query sequence indeed had been reverse complementary initially. Of the 1,000 cases where the program could not find the 5.8S, resulting in no decision on orientation being reached, 870 (87%) could be explained by the absence—or presence of <50 bp—of the 5.8S; 78 (7.8%) represented taxonomic lineages with very divergent ribosomal sequences (below); 41 (4.1%) had sequence data of poor read quality (e.g., high content of ambiguity symbols) in the 5.8S region targeted; and 11 (1.1%) were of non-fungal origin. No clear false negatives or false positives were thus found, and we conclude that the performance of the software is as good as can be expected given the heterogeneous nature of the international corpus of fungal sequence data.

The 161,922-sequence dataset was analyzed in 57 min, with the option to oversee BLAST searches and MAFFT alignments deactivated, on a dual-core, 2.2-GHz MacBook Pro laptop. The BLAST/MAFFT option slowed down the analyses considerably, with 1,000 random INSD entries analyzed in 75 min. An obvious shortcoming of the present software is the requirement that approximately 50 bp of the 5′-end of 5.8S be present in the query sequences for them to be fully processed. This restriction rules out many partial ITS sequences, including those from several massively parallel (454) pyrosequencing studies, from being analyzed by the software. (The software does, however, support the inclusion of a second set of HMMs to cover the less-conserved 3′-end of 5.8S.) Taxa with very deviant ribosomal genes, such as *Cantharellus* and *Tulasnella* (Feibelman et al. 1994; Moncalvo et al. 2006; Taylor and McCormick 2008), may have 5.8S sequences different enough to escape detection by the HMMs employed in the present software; indeed, *Tulasnella* alone accounted for more than 70% of the 78 cases where the 5.8S was present but remained undetected. These taxa tend not to be picked up by any primer combination regularly used in environmental sequencing. Therefore, any user expecting to analyze such deviant lineages should consider constructing separate HMMs for these; expanding general fungal HMMs to also cover these lineages may detract from the performance of the HMMs on the majority of fungi. The default settings of the software package are stringently set to keep the number of false positives at an absolute minimum. These settings include making use of a comparatively long stretch of the 5′-end of the 5.8S and HMMER *E*-values tailored to exclude spurious matches. HMMs for several other groups of organisms are provided to facilitate the implementation of the software for non-fungal organisms (e.g., plants, animals, and oomycetes; Supplementary Item 3).

We found that 1,443 of 161,922 (0.9%) reasonably full-length fungal ITS sequences in INSD were given in the reverse complementary orientation by the original sequence authors. About half of these entries were older than 5 years and some older than 10 years. Although new features are regularly added to the INSD sequence analysis toolbox to improve the integrity of the data, including a reverse complementary checker for the small-subunit (16S/18S) ribosomal gene, the foregoing observations stress the need for critical examination of publicly available sequence data before inclusion in scientific analyses (Harris 2003). Interestingly, the majority (62%) of the reverse complementary entries were not identified to species level, suggesting that they are the result of environmental sequencing efforts and perhaps automated sequence generation pipelines rather than manual assembly and inspection. The number of environmental DNA sequences is expected to increase dramatically in the wake of

emerging sequencing technologies (Hibbett et al. 2009), and there is every reason to be careful when such datasets are analyzed and used by other researchers. The INSD policies for third-party updating and annotation of INSD entries are presently restrictive enough that such interactions with the entries are reserved for the original sequence authors only (Pennisi 2008), although improvements may be looming on the horizon. The fungal ITS sequences in INSD are, however, mirrored by the UNITE database for molecular identification of fungi (Abarenkov et al. 2010; http://unite.ut.ee), a database that supports third-party annotation of sequence data. All 1,443 entries found to be incorrectly oriented by the present effort were thus annotated in UNITE to alert the user to the fact that, as presently given in INSD, these entries are reverse complementary. We used the software to reorient these entries such that they no longer pose a potential source of artificial signal to anyone accessing INSD data through UNITE. We similarly reported them to INSD such that they are likely to be reoriented there also in the foreseeable future. Even so, the constant stream of sequence submissions suggests that the issue of reverse complementary fungal ITS sequences in public sequence repositories will not be resolved until either the sequence submitter or the receiving party, and preferably both, take measures to counter such entries. We view the present software as a small step in that direction, and one that comes at no cost and a very modest time consumption for the user.

Acknowledgments R.H.N. and K.A. gratefully acknowledge support from the Frontiers in Biodiversity Research Centre of Excellence (University of Tartu) and the Fungi in Boreal Forest Soils network. Matt von Konrat and Anders Hagborg are acknowledged for valuable advice on the liverwort data. Two anonymous reviewers are acknowledged for valuable input on the manuscript. The authors declare that they have no conflict of interests. No laboratory experiments were undertaken as a component—or result—of the present study.

References

Abarenkov K, Nilsson RH, Larsson K-H et al (2010) The UNITE database for molecular identification of fungi: recent updates and future perspectives. New Phytol 186:281–285

Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Begerow D, Nilsson RH, Unterseher M, Maier W (2010) Current state and perspectives of fungal DNA barcoding and rapid identification procedures. Appl Microbiol Biotechnol 87:99–108

Bidartondo MI, Bruns TD, Blackwell M et al (2008) Preserving accuracy in GenBank. Science 319:1616

Blackwell M, Hibbett DS, Taylor JW, Spatafora JW (2006) Research coordination networks: a phylogeny for kingdom Fungi (Deep Hypha). Mycologia 98:829–837

Eckart M, Fliegerova K, Hoffmann K, Voigt K (2010) Molecular identification of anaerobic rumen fungi. In: Gherbawy Y, Voigt K (eds) Molecular identification of fungi. Springer, New York, pp 297–313

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763

Feibelman TP, Bayman P, Cibula WG (1994) Length variation in the internal transcribed spacer of ribosomal DNA in chanterelles. Mycol Res 98:614–618

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221–224

Harris DJ (2003) Can you bank on GenBank? Trends Ecol Evol 18:317–319

Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH (2010) V-Xtractor: an open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16 S/18 S) ribosomal RNA gene sequences. J Microbiol Methods 83:250–253

Hibbett DS (2007) After the gold rush, or before the flood? Evolutionary morphology of mushroom-forming fungi (*Agaricomycetes*) in the early 21st century. Mycol Res 111:1001–1018

Hibbett DS, Nilsson RH, Snyder M, Fonseca M, Costanzo J, Shonfeld M (2005) Automated phylogenetic taxonomy: an example in the homobasidiomycetes. Syst Biol 54:660–668

Hibbett DS, Ohman A, Kirk PM (2009) Fungal ecology catches fire. New Phytol 184:279–282

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286–298

Kauff F, Cox C, Lutzoni F (2007) WASABI: an automated sequence processing system for multi-gene phylogenies. Syst Biol 56:523–531

Moncalvo J-M, Nilsson RH, Koster B et al (2006) The cantharelloid clade: dealing with incongruent gene trees and phylogenetic reconstruction methods. Mycologia 98:937–948

Nilsson RH, Kristiansson E, Ryberg M, Larsson K-H (2005) Approaching the taxonomic affiliation of unidentified sequences in public databases: an example from the mycorrhizal fungi. BMC Bioinform 6:178

Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson K-H (2008) Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. Evol Bioinf Online 4:193–201

Nilsson RH, Abarenkov K, Veldre V, Nylinder S, De Wit P, Brosché S, Alfredsson JF, Ryberg M, Kristiansson E (2010) An open source chimera checker for the fungal ITS region. Mol Ecol Res 10:1076–1081

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

Pennisi E (2008) Proposal to 'Wikify' GenBank meets stiff resistance. Science 319:1598–1599

Porter TM, Skillman JE, Moncalvo J-M (2008) Fruiting body and soil rDNA sampling detects complementary assemblage of *Agaricomycotina* (*Basidiomycota*, *Fungi*) in a hemlock-dominated forest plot in southern Ontario. Mol Ecol 17:3037–3050

Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH (2009) An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity. New Phytol 181:471–477

Sayers EW, Barrett T, Benson DA et al (2010) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 38:D5–D16

Seifert KA (2009) Progress towards DNA barcoding of fungi. Mol Ecol Res 9:83–89

Stajich JE, Block D, Boulez K et al (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12:1611–1618

Suzuki A, Bärlocher F (2009) Editorial for the special feature: propagation strategy of fungi. Mycoscience 50:1–2

Taylor DL, McCormick K (2008) Internal transcribed spacer primers and sequences for improved characterization of basidiomycetous orchid mycorrhizas. New Phytol 177:1020–1033

Taylor JW, Jacobson DJ, Kroken S, Kasuga T, Geiser DM, Hibbett DS, Fisher MC (2000) Phylogenetics species recognition and species concepts in fungi. Fungal Genet Biol 31:21–32