Ferro et al. BMC Research Notes 2014, 7:857
http://www.biomedcentral.com/1756-0500/7/857

BMC Research Notes

TECHNICAL NOTE Open Access

ITScan: a web-based analysis tool for Internal Transcribed Spacer (ITS) sequences

Milene Ferro[1*], Erik A Antonio[3], Wélliton Souza[1] and Maurício Bacci Jr[1,2]

Abstract

Background: Studies on fungal diversity and ecology aim to identify fungi and to investigate their interactions with each other and with the environment. DNA sequence-based tools are essential for these studies because they can speed up the identification process and access greater fungal diversity than traditional methods. The nucleotide sequence encoding for the internal transcribed spacer (ITS) of the nuclear ribosomal RNA has recently been proposed as a standard marker for molecular identification of fungi and evaluation of fungal diversity. However, the analysis of large sets of ITS sequences involves many programs and steps, which makes this task intensive and laborious.

Findings: We developed the web-based pipeline ITScan, which automates the analysis of fungal ITS sequences generated either by Sanger or Next Generation Sequencing (NGS) platforms. Validation was performed using datasets containing ca. 2,000 to 40,000 sequences each.

Conclusions: ITScan is an online and user-friendly automated pipeline for fungal diversity analysis and identification based on ITS sequences. It speeds up a process which would otherwise be repetitive and time-consuming for users. The ITScan tool and documentation are available at http://evol.rc.unesp.br:8083/itscan.

Keywords: Fungal biodiversity, Mycology, Pipeline, Web service

Findings

Background

Studies on fungal biodiversity use DNA sequence-based tools to generate molecular marker to identify rare species and determine associations in a microbial community [1]. The technique is particularly powerful in characterizing fungal diversity in environmental samples containing many fungal species which do not grow, or grow poorly, in laboratory cultures [2]. Many biodiversity studies are based on the nuclear ribosomal Internal Transcribed Spacer (ITS) region [3,4], which is a small (~500 base-pair) region occurring in multiple copies in the fungal nuclear genome and shows a high degree of variation even between closely related species [5].

The ITS region has been recently designated as a universal marker for molecular barcoding of fungi [1] or the default region for species identification. To determine the microbial diversity in environmental samples, generated ITS sequences are grouped in operational taxonomic units (OTUs), often using the MOTHUR program [6] and an OTU-based approach analysis [7,8]. The use of multiple programs and stages of analysis make the process laborious and time-consuming. In this work, we describe a web-based pipeline that automates the study of fungal diversity and identification based on ITS sequences.

Implementation

Architecture design

We developed an architectural model based on MVC (Model-View-Controller) and J2EE design patterns [9] (Figure 1). The architectural model also depicts two base formats for data interchange: JavaScript Object Notation (JSON) and Extensible Markup Language (XML). These formats represent data and functions as well as each step used in the pipeline architecture to perform fungal analysis. The architecture model was tailored to represent two main viewpoints:

* Correspondence: milenef@gmail.com
[1]Centro de Estudos de Insetos Sociais, Instituto de Biociências, UNESP - Univ Estadual Paulista, Rio Claro SP 13506-900, Brazil
Full list of author information is available at the end of the article

BioMed Central

© 2014 Ferro et al.; 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/4.0), 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.

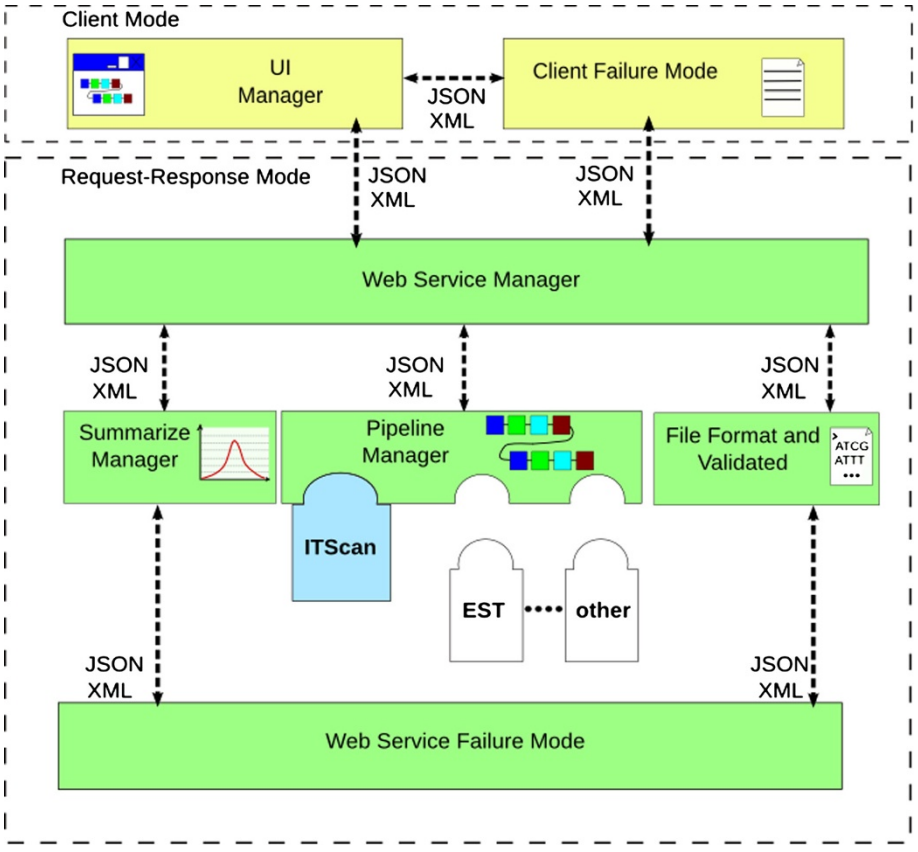

Figure 1 System architecture that coupled ITScan. The figure displays the ITScan architecture model based on MVC (Model-View-Controller) and J2EE design patterns. The architecture model was tailored to represent two main viewpoints: Client Mode and Request-Response Mode.

- Client Mode — aims at dealing with client-side concerns;
- Request-Response Mode — performs a set of server-side and business logic concerns using coupled third-party programs and their business rules. The Pipeline Manager provides Representation State Transfer - REST [10] service.

This architecture assists background information to check for failures in client and server sides.

Pipeline for fungal ITS analysis

ITScan requires a FASTA-formatted input file containing pre-processed sequences, i.e., high quality sequences (usually Phred ≥20) without primer and adaptor sequences. Pre-processing programs, such as SEQTRIM [11], SCATA [12], PANGEA [13], CANGS [14] and PYRONOISE [15], can be used to trim data from different sequencing platforms (e.g. 454, Illumina, regular Sanger reads) and the resulting output files can then be read by ITScan.

The third-party programs ChimeraChecker [16], MAFFT [17], MOTHUR and BLAST [18] were integrated in the pipeline as shown by the state machine diagram using UML [19] (Figure 2). Each program in ITScan is a web service developed using REST technology, which was shown to improve client usability [20,21]. In the first step, ChimeraChecker is used to classify all sequences as chimeric, non-chimeric or not evaluated using default parameters. Non-chimeric ITS sequences are then aligned to each other in the MAFFT software. Aligned sequences are run into the MOTHUR package, which clusters similar sequences to each other to generate operational taxonomic units (OTUs), and calculates diversity indexes and richness estimators [6]. User can set the ITScan label parameter to define the dissimilarity value (%) that represents the maximal percentage of difference between the sequences in the same OTU. MOTHUR selects a representative sequence which has the smallest distance from all remaining sequences within a given OTU. The selected representative sequence (or centroid) is used in a BLASTN search and the first hit is used to identify the OTU. The utilization of a centroid instead of all sequences composing the OTU speeds up computation processing. BLAST results are presented in tabular format with links to GenBank.

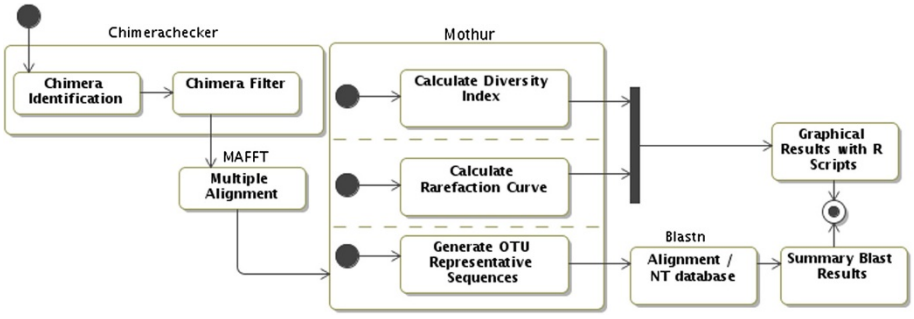

Figure 2 State machine diagram describing ITScan pipeline steps. The third-party programs were integrated in the pipeline as shown by the state machine diagram using UML. Each program in ITScan is a web service developed using REST.

Results

The architectural model enables the user to develop web service components and to couple them in a new customized pipeline. R language scripts provide graphic results and spreadsheets representing rarefaction curves as well as Shannon or Simpson diversity indexes and Chao1 richness estimator.

ITScan has a user-friendly interface and can process up to three a FASTA-formatted input files simultaneously and compare these files with each other. The pipeline was validated using Sanger sequences (Mantovani et al., in preparation) and a large dataset (2,000 to 40,000 sequences) simulating results from Next Generation Sequencing (NGS), which was retrieved from the UNITE [22] database.

Many programs which analyze ITS fungal sequences, such as FungalITSPipeline [23], QIIME [24] and FHiTINGS [25], require the user installation and operation via command line. These requirements are not necessary in ITScan, which was built with a web-based interface.

The ITScan pipeline comes with some limitations. For instance, it processes only three FASTA files simultaneously. In addition, it relies on GenBank servers to run BLASTN searches, instead of implementing time-consuming local searches on annotated databases [22] which would improve taxonomic assignment. Future expansions in our servers will allow us to implement multi sample analyses based on local annotated fungal ITS databases.

Conclusions

This work describes an architectural model that can be used with bioinformatics third-party programs. All components follow the same framework, which facilitates the development of new components. ITScan works with sequences derived from both Sanger and NGS technologies. The pipeline can process single or as many as three datasets to compare distinct biological samples. Output data include graphs and spreadsheets that are automatically generated to represent fungal diversity. ITScan includes an user manual and an example dataset. We validated ITScan using datasets containing ca. 2,000 and 40,000 sequences retrieved from the UNITE database. Using of ITScan does not require computational expertise.

Availability and requirements

Project name: ITScan
Project home page: http://evol.rc.unesp.br:8083/itscan
Operating system(s): Platform independent
Programming language: Perl, Java
Other requirements: Web browser
License: ITScan web tool is freely available for all users. ITScan is open source under the GNU GPL license.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Pipeline conceiving: MF and MB. Pipeline components development: MF, WS and EAA. Architecture design: EAA and MF. Pipeline validation: MF. Manuscript preparation and revision: MF, EAA and MB. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (Proc. 2011/50226-0). MF receives a doctoral fellowship (Proc. 2009/52289-9).

Author details

[1]Centro de Estudos de Insetos Sociais, Instituto de Biociências, UNESP - Univ Estadual Paulista, Rio Claro SP 13506-900, Brazil. [2]Departamento de Bioquímica e Microbiologia, Instituto de Biociências, UNESP - Univ Estadual Paulista, Rio Claro, SP 13506-900, Brazil. [3]Departamento de Ciência da Computação, Universidade Federal de São Carlos, São Carlos, SP 13565-905, Brazil.

Received: 18 March 2014 Accepted: 19 November 2014
Published: 27 November 2014

References

1. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Fungal Barcoding Consortium: **Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi.** *Proc Natl Acad Sci* 2012, **109**:6241–6246.
2. Sun X, Guo LD: **Endophytic fungal diversity: review of traditional and molecular techniques.** *Mycology* 2012, **3**(1):65–76.

3. Rittenour WR, Ciaccio CE, Barnes CS, Kashon ML, Lemons AR, Beezhold DH, Green BJ: **Internal transcribed spacer rRNA gene sequencing analysis of fungal diversity in Kansas City indoor environments.** *Environ Sci Process Impacts* 2014, **16**(1):33–43.
4. Liu YT, Chen RK, Lin SJ, Chen YC, Chin SW, Chen FC, Lee CY: **Analysis of sequence diversity through internal transcribed spacers and simple sequence repeats to identify Dendrobium species.** *Genet Mol Res* 2014, **13**(2):2709–2717.
5. Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H: **ITS as an environmental DNA barcode for fungi: an in silico approach reveals PCR biases.** *BMC Microbiol* 2010, **10:**189.
6. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF: **Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities.** *Appl Environ Microbiol* 2009, **75**(23):7537–7541.
7. Schloss PD, Westcott SL: **Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis.** *Appl Environ Microbiol* 2011, **77**(10):3219–3226.
8. Links MG, Chaban B, Hemmingsen SM, Muirhead K, Hill JE: **mPUMA: a computational approach to microbiota analysis by de novo assembly of operational taxonomic units based on protein-coding barcode sequences.** *Microbiome* 2013, **1:**23.
9. Fowler M: *Patterns of Enterprise Application Architecture.* Boston: Addison-Wesley; 2002.
10. Richardson L, Ruby S: *RESTful Web Services: Web Services for the Real World.* Sebastopol: O'Reilly Media; 2007.
11. Falgueras J, Lara AJ, Fernández Pozo N, Cantón FR, Pérez Trabado G, Claros MG: **SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read.** *BMC Bioinform* 2010, **20**(11):38.
12. SCATA: **Sequence Clustering and Analysis of Tagged Amplicons.** In http://scata.mykopat.slu.se/.
13. Giongo A, Crabb DB, Davis-Richardson AG, Chauliac D, Mobberley JM, Gano KA, Mukherjee N, Casella G, Roesch LF, Walts B, Riva A, King G, Triplett EW: **PANGEA: pipeline for analysis of next generation amplicons.** *ISME J* 2010, **4**(7):852–861.
14. Pandey RV, Nolte V, Schlotterer C: **CANGS: a user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies.** *BMC Res Notes* 2010, **11**(3):3.
15. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT: **Accurate determination of microbial diversity from 454 pyrosequencing data.** *Nat Methods* 2009, **6**(9):639–641.
16. Nilsson RH, Abarenkov K, Veldre V, Nylinder S, Wit P, Brosché S, Alfredsson JF, Ryberg M, Kristiansson E: **An open source chimera checker for the fungal ITS region.** *Mol Ecol Resour* 2010, **10:**1076–1081.
17. Katoh M, Kuma M: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res* 2002, **30:**3059–3066.
18. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–3402. 1.
19. Booch G, Rumbaugh J, Jacobson I: *Unified Modeling Language User Guide.* Reading: Addison-Wesley Professional; 2005.
20. Katayama T, Nakao M, Takagi T: **TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services.** *Nucleic Acids Res* 2010, **38:** W706–W711.
21. Medina I, De Maria A, Bleda M, Salavert F, Alonso R, Gonzalez CY, Dopazo J: **VARIANT: command Line, web service and web interface for fast and accurate functional characterization of variants found by next-generation sequencing.** *Nucleic Acids Res* 2012, **40:**W40–W58.
22. Abarenkov K, Nilsson RH, Larsson K, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Ursing BM, Vrålstad T, Liimatainen K, Peintner U, Kõljalg U: **The UNITE database for molecular identification of fungi - recent updates and future perspectives.** *New Phytol* 2010, **186**(2):281–285.
23. Nilsson RH, Bok G, Ryberg M, Kristiansson E, Hallenberg N: **A software pipeline for processing and identification of fungal ITS sequences.** *Source Code Biol Med* 2009, **4:**1.
24. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R: **QIIME allows analysis of high-throughput community sequencing data.** *Nat Methods* 2010, **7**(5):335–336.
25. Dannemiller KC, Reeves D, Bibby K, Yamamoto N, Peccia J: **Fungal high-throughput taxonomic identification tool for use with next-generation sequencing (FHiTINGS).** *J Basic Microbiol* 2014, **54:**315–321.

doi:10.1186/1756-0500-7-857
Cite this article as: Ferro *et al.*: ITScan: a web-based analysis tool for Internal Transcribed Spacer (ITS) sequences. *BMC Research Notes* 2014 **7**:857.

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