Crowth, an identification platform for grapevine, olive and honey

Antonios Kioukis,[1] Ilias Lagkoubardos,[2] and Pavlos Pavlidis[3]

[1] *University of Crete, Medical School*
[2] *Technical University Munich — TUM ZIEL - Institute for Food and Health*
[3] *Foundation for Research and Technology - Hellas, Institute of Computer Science*

The Internal Transcriber Spacer (ITS) region has been proposed to act as the universal DNA barcode for plants. Here, we present Crowth (CRetan grOWTH), a web platform that identifies and quantifies the plant origins of three Cretan products by creating a genetic identity based on their ITS region. Furthermore, each sequence of interest is placed in a phylogenetic tree to allow for broader evidences of similarity. To our best knowledge, Crowth is the first web server dedicated to the identification and quantification of wine, olive oil and honey using the ITS region, and currently hosts more than two hundred plants endemic in Crete. Crowth is available at *http:139.91.68.81/*

I. INTRODUCTION

Internal Transcriber Spacers (ITS1, ITS2) are spacer DNA located between the small-subunit ribosomal RNA (rRNA) and large-subunit rRNA genes. In plants, ITS1 is located between 18S and 5.8S rRNA genes, while ITS2 is between 5.8S and 26S. ITS1 and ITS2 have long been used as a region for phylogenetic reconstruction of species and genus relationships [1–4] using comparisons of primary sequence. The usage of ITS makes possible the creation of reliable sequence-structure alignments that take into account the secondary structure of the region due to its high conservation within all eukariotes [5–7]. The comparison of sequences based on the ITS region is widely used in taxonomy [8] and molecular phylogeny because of several favorable properties. Its small size allows for amplification and association with available highly conserved flanking sequences. It is detectable even from small quantities of DNA due to the high copy number of the rRNA clusters [9]. Unequal crossing-over and gene conversion result in rapid concerted evolution. This promotes intra-genomic homogeneity of the repeat units, although high-throughput sequencing showed the occurrence of frequent variations within plant species. Finally, it has a high degree of variation even between closely related species. This can be explained by the relatively low evolutionary pressure acting on such non-coding spacer sequences. This conservation permits comparisons at deeper taxonomic levels [10–16]. Based on these facts we created Crowth , a platform for the genetic identification of three Cretan products wine, olive oil and honey. Crowth is based on the Internal Transcriber Spacers(ITS1, ITS2) to create genetic identities for each of the plants. The genetic identity is more specific in grapevine and olive trees differentiating between different cultivars where as the genetic identity of flowers signals a higher taxonomic level. The diversity of Cretan micro-climates combined with the island's altitude differences enhance the diversification of flora and allows for plants of different taxonomic groups to co-exist and mix. Crowth provides the necessary framework to get back at the source of each product and identify whence it came from.

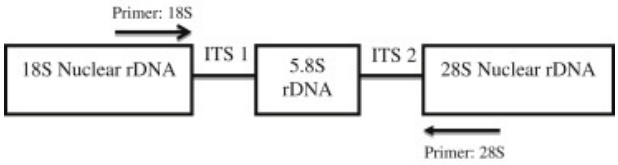

FIG. 1. ITS locations between 18S,5.8S and 26S genes

II. MATERIALS AND METHODS

A. Database Schema

The core of Crowth is comprised of three tables (Grapevine, Flower and Olive) that contain each products' available information. For every plant in the database Crowth stores: (i) a unique integer identifier used as the primary key of the table. (ii) The name of the plant originated from the NCBI downloaded file. (iii) The ITS sequence is stored in the sequence field of the table. (iv) The last updated field holds information showing when the table entry was last modified. (v) Cultivar description is also taken from the NCBI file and holds all the information besides name and sequence provided by the NCBI file. (vi) The link field is currently empty but when populated will provide a hypertext link to a page holding general information about the plant in question.

Crowth operates on two categories of queries. Identification Queries accurately predict the closest taxa from the user-provided sequences. Quantification Queries handle metagenomic samples, by processing a FASTQ file containing amplicons from a PCR experiment. The results of all queries are available for download from the main dashboard located in /jobs.html.

B. Query Analysis Overview

Identification Queries

Identification Queries are further divided in two categories: Simple sequence repeats (SSRs) and ITS, depending on what region will be used for the identification

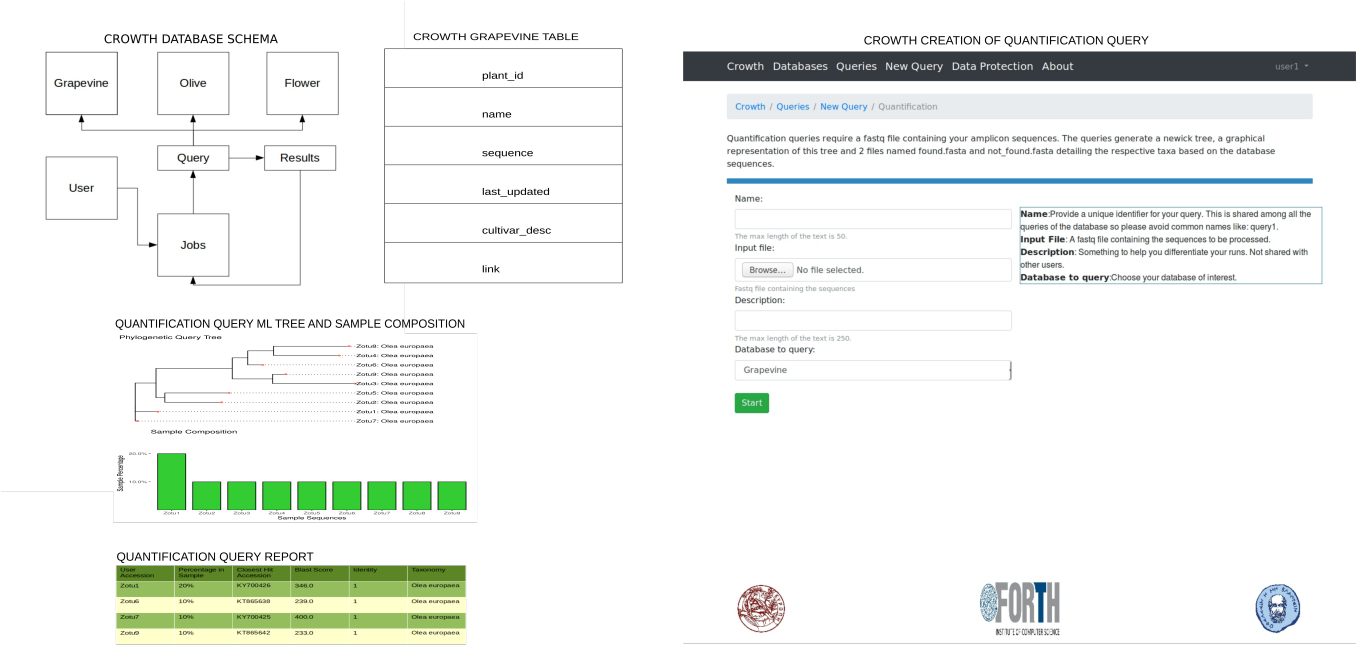

FIG. 2. Top left: Crowth database schema, showing the process of each new query. Top Middle: Grapevine table field. Right: Query creation sceen as viewed by the front-end. Bottom left: Quantification Query results available for download.

process.

SSRs analysis is composed of two phases, parsing and distance calculation. Crowth currently supports 18 different SSR locations. The first step of each SSR analysis is to identify which of the 18 markers currently supported by Crowth are contained in the input. Locations not present in the input file do not affect the analysis results. It is worth noting that the robustness of the analysis and the confidence in the results are directly correlated with the number of included SSR locations. Crowth calculates the distance between the provided sample and the locations in the database. The distance metric used is the euclidean distance due to the linearity of the data and its low processing resources requirements. The SSR identification analysis is currently available only for grapevine.

The identification analysis based on the ITS region requires more steps. The input sequences are BLASTn [17] against the target products' database sequences. The top N hits for each sequence are saved in a tab-delimited file. Next, the input sequences are used to create a maximum likelihood phylogenetic tree [18]. To do that, the input sequences must be aligned with the N best hit database sequences. For this purpose we use the MAFFT aligner [19]. The alignment of database and query sequences is calculated from scratch everytime because Crowth has no knowledge a priori of the database sequences that would be the top hits. After the new alignment is done, the phylogenetic tree is created using RAxML [18] and is stored in newick format. The resulting tree is used for a preliminary visualization through the use of a custom R script.

Quantification Queries

Quantification queries are handled by two already developed tools Usearch [20] and RAxML [18]. The first step is dereplication which finds the set of unique sequences from the FASTQ file. Dereplication compares all the sequences from the input file and extracts sequences that match exactly. Denoising [21] is the next step, sequence errors from amplicon reads are removed and the correct biological sequences in the reads are identified. The output sequences are now devoid of errors and are placed in a FASTA file, followed by a BlastN search against the Crowth database. Sequences that match with the database are excluded from further analysis and are placed in a file available for download when the query has completed. Sequences that are not matched are placed in the database's phylogenetic tree to identify the closer taxa. This action offers additional information for the sample.

C. Implementation

Crowth is implemented using the Django python framework. Django implements a MVC (Model-View-Controller) architecture, consisting of an object-relational mapper that interacts with data models in the relational database ("Model"), a handler of HTTP re-

quests with a web templating system ("View"), and a regular-expression-based URL dispatcher ("Controller"). Crowth currently supports up to 3 concurrent queries independent of the query type. Crowth guarantees to maintain query results for at least two weeks. The actual processing of the data is handled by custom made python scripts implementing a many to one access to the Crowth database. Each query is scheduled with the use of Celery [**?**]. Celery is an open-source asynchronous task queue or job queue which is based on distributed message passing. NGINX is used as the back-end HTTP server and reverse proxy. NGINX was chosen because it does not rely on threads to handle requests like traditional server (apache2). Instead it uses a much more scalable event-driven (asynchronous) architecture which uses small, but more importantly, predictable amounts of memory under load. Supervisor is used as a fail-safe to automatically detect anomalies such as the Django back-end or the NGINX front-end are shut down, to restart them to lose the minimal response time.

D. Data Export and Search

Crowth enables each user to export the database data. Each database has a dedicated link to generate a file containing all the respective information. This file is either in comma separated format or fasta. Searches on the products database can be conducted using as filter either the NCBI accession name or a part of the description. The search results are also available for download.

III. SERVER DESCRIPTION

Crowth is currently available at: http://139.91.68.4:8080/.

A. Input

Crowth can be queried using a DNA sequences in the FASTA format for identification queries. Quantification queries require a FASTQ file of the query sequences. The length of the input sequences is not limited and the time required for identification queries is typically very fast (¡1 min). However, it may take up to several minutes for big inputs. Quantification queries, typically, take longer. A bare bones programming interface is currently being developed.

B. Output

Crowth provides output for all the three type of queries on the same page. This output is available for download for two weeks after each query has completed. Each query output bundles together: a general report explaining in detail how the output files where generated and what they report, the specific tab-delimited query report file, the phylogenetic trees and their visualizations.

IV. CONCLUSION AND FUTURE PERSPECTIVES

Crowth is a robust solution for identification and quantification for every producer and consumer based on NGS technologies. With the inclusion of SSR methods for backwards compatibility, Crowth seeks to extend the use of NGS identification methods on grapevine, olive oil and honey without alienating current approaches. In the future Crowth will include a picture of each plant as well as geographical links of its known habitats.

[1] Hui Yao, Jingyuan Song, Chang Liu, Kun Luo, Jianping Han, Ying Li, Xiaohui Pang, Hongxi Xu, Yingjie Zhu, Peigen Xiao, and Shilin Chen. Use of its2 region as the universal dna barcode for plants and animals. *PLOS ONE*, 5(10):1–9, 10 2010.

[2] Annette W Coleman. Its2 is a double-edged tool for eukaryote evolutionary comparisons. *TRENDS in Genetics*, 19(7):370–375, 2003.

[3] Annette W Coleman. Pan-eukaryote its2 homologies revealed by rna secondary structure. *Nucleic Acids Research*, 35(10):3322–3329, 2007.

[4] Annette W Coleman. Is there a molecular key to the level of biological species in eukaryotes? a dna guide. *Molecular Phylogenetics and Evolution*, 50(1):197–203, 2009.

[5] Jörg Schultz, Stefanie Maisel, Daniel Gerlach, Tobias Müller, and Matthias Wolf. A common core of secondary structure of the internal transcribed spacer 2 (its2) throughout the eukaryota. *Rna*, 11(4):361–364, 2005.

[6] Jörg Schultz, Tobias Müller, Marco Achtziger, Philipp N Seibel, Thomas Dandekar, and Matthias Wolf. The internal transcribed spacer 2 databasea web server for (not only) low level phylogenetic analyses. *Nucleic Acids Research*, 34(suppl_2):W704–W707, 2006.

[7] Jörg Schultz and Matthias Wolf. Its2 sequence–structure analysis in phylogenetics: a how-to manual for molecular systematics. *Molecular Phylogenetics and Evolution*, 52(2):520–523, 2009.

[8] Hui Yao, Jingyuan Song, Chang Liu, Kun Luo, Jianping Han, Ying Li, Xiaohui Pang, Hongxi Xu, Yingjie Zhu, Peigen Xiao, et al. Use of its2 region as the universal dna barcode for plants and animals. *PloS one*, 5(10):e13102,

2010.

[9] Jingyuan Song, Linchun Shi, Dezhu Li, Yongzhen Sun, Yunyun Niu, Zhiduan Chen, Hongmei Luo, Xiaohui Pang, Zhiying Sun, Chang Liu, Aiping Lv, Youping Deng, Zachary Larson-Rabin, Mike Wilkinson, and Shilin Chen. Extensive pyrosequencing reveals frequent intra-genomic variations of internal transcribed spacer regions of nuclear ribosomal dna. *PLOS ONE*, 7(8):1–12, 08 2012.

[10] Shilin Chen, Hui Yao, Jianping Han, Chang Liu, Jingyuan Song, Linchun Shi, Yingjie Zhu, Xinye Ma, Ting Gao, Xiaohui Pang, et al. Validation of the its2 region as a novel dna barcode for identifying medicinal plant species. *PloS one*, 5(1):e8613, 2010.

[11] Ting Gao, Hui Yao, Jingyuan Song, Chang Liu, Yingjie Zhu, Xinye Ma, Xiaohui Pang, Hongxi Xu, and Shilin Chen. Identification of medicinal plants in the family fabaceae using a potential dna barcode its2. *Journal of ethnopharmacology*, 130(1):116–121, 2010.

[12] Xiaohui Pang, Jingyuan Song, Yingjie Zhu, Hongxi Xu, Linfang Huang, and Shilin Chen. Applying plant dna barcodes for rosaceae species identification. *Cladistics*, 27(2):165–170, 2011.

[13] Kun Luo, ShiLin Chen, KeLi Chen, JingYuan Song, Hui Yao, Xinye Ma, YingJie Zhu, XiaoHui Pang, Hua Yu, XiWen Li, et al. Assessment of candidate plant dna barcodes using the rutaceae family. *Science China Life Sciences*, 53(6):701–708, 2010.

[14] Yanwei Li, XIN Zhou, Gui Feng, HaoYuan Hu, LiMing Niu, Paul DN Hebert, and DaWei Huang. Coi and its2 sequences delimit species, reveal cryptic taxa and host specificity of fig-associated sycophila (hymenoptera, eurytomidae). *Molecular ecology resources*, 10(1):31–40, 2010.

[15] Pramod Kumar Prasad, Veena Tandon, Devendra Kumar Biswal, Lalit Mohan Goswami, and Anupam Chatterjee. Phylogenetic reconstruction using secondary structures and sequence motifs of its2 rdna of paragonimus westermani (kerbert, 1878) braun, 1899 (digenea: Paragonimidae) and related species. In *BMC genomics*, volume 10, page S25. BioMed Central, 2009.

[16] Pramod Kumar Prasad, Veena Tandon, Devendra Kumar Biswal, Lalit Mohan Goswami, and Anupam Chatterjee. Use of sequence motifs as barcodes and secondary structures of internal transcribed spacer 2 (its2, rdna) for identification of the indian liver fluke, fasciola (trematoda: Fasciolidae). *Bioinformation*, 3(7):314, 2009.

[17] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1):421, 2009.

[18] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[19] Kazutaka Katoh and Daron M. Standley. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

[20] Robert C Edgar. Unoise2: improved error-correction for illumina 16s and its amplicon sequencing. *bioRxiv*, 2016.

[21] Robert C. Edgar and Henrik Flyvbjerg. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, 31(21):3476–3482, 2015.