Neotrop Entomol (2016) 45:554–558
DOI 10.1007/s13744-016-0403-0

SYSTEMATICS, MORPHOLOGY AND PHYSIOLOGY

Toward an Automated Identification of *Anastrepha* Fruit Flies in the *fraterculus* group (Diptera, Tephritidae)

P Perre[1], FA Faria[2], LR Jorge[3], A Rocha[4], RS Torres[4], MF Souza-Filho[5], TM Lewinsohn[3], RA Zucchi[1]

[1]Dept of Genectis and Evolutionary Biology, Univ of São Paulo, São Paulo, SP, Brasil
[2]Institute of Science and Technology, Federal Univ of São Paulo, São José dos Campos, SP, Brasil
[3]Dept of Animal Biology, Institute of Biology, Univ of Campinas, Campinas, SP, Brasil
[4]Institute of Computing, Univ of Campinas, Campinas, SP, Brasil
[5]Economic Entomology Lab, Biological Institute, APTA, Campinas, SP, Brasil

Keywords
Identification, image analysis, machine learning

Correspondence
P Perre, Dept of Genetics and Evolutionary Biology, IB, Univ of São Paulo, Rua do Matão 277, São Paulo 05508-090, SP, Brasil; pperre01@gmail.com

Edited by Ranyse B Querino – Embrapa

Received 19 January 2016 and accepted 6 April 2016
Published online: 8 May 2016

Abstract
In this study, we assess image analysis techniques as automatic identifiers of three *Anastrepha* species of quarantine importance, *Anastrepha fraterculus* (Wiedemann), *Anastrepha obliqua* (Macquart), and *Anastrepha sororcula* Zucchi, based on wing and aculeus images. The right wing and aculeus of 100 individuals of each species were mounted on microscope slides, and images were captured with a stereomicroscope and light microscope. For wing image analysis, we used the color descriptor *Local Color Histogram*; for aculei, we used the contour descriptor *Edge Orientation Autocorrelogram*. A *Support Vector Machine* classifier was used in the final stage of wing and aculeus classification. Very accurate species identifications were obtained based on wing and aculeus images, with average accuracies of 94 and 95%, respectively. These results are comparable to previous identification results based on morphometric techniques and to the results achieved by experienced entomologists. Wing and aculeus images produced equally accurate classifications, greatly facilitating the identification of these species. The proposed technique is therefore a promising option for separating these three closely related species in the *fraterculus* group.

Introduction

Taxonomy is essential to many important applications in biology (Bortolus 2008). Classic morphology-based taxonomy—performed mainly by specialized taxonomists—has been failing to meet the demand for species identification, characterization, classification, and description (Weeks & Gaston 1997). Taxonomy has thus turned into a shortfall in many branches of biology, limiting their development (Wheeler *et al* 2004, Hortal *et al* 2015). A possible means of overcoming this taxonomic impediment is to design and develop alternatives that non-taxonomists can use in their studies, without complete dependence on the availability of specialists (Godfray 2002, MacLeod *et al* 2010).

Many studies in the literature have used image processing in the pursuit of solutions to identification problems by way of machine learning. A number of alternatives have been tested, including genetic identification by molecular barcoding (Hebert *et al* 2003, Hebert & Gregory 2005); online access to simplified identification keys with detailed images of important structures for species identification (Gaston & O'Neil 2004); morphometric techniques, multivariate and/or geometric (Baylac *et al* 2003); and computer-based identification via image analysis (Weeks *et al* 1999). The strongest motivation for pursuing these techniques is their ability to automate the identification process.

One of the advantages of automating the species identification process by means of image analysis and

machine learning is that it reduces dependence on expert opinions. This makes taxonomy accessible to all researchers who need to recognize specimens, whether these researchers are experts or not (LaSalle *et al* 2009, MacLeod *et al* 2010). Notable publications have already presented alternatives for insect identification (Arbuckle *et al* 2001, Watson *et al* 2003, Russell *et al* 2007, Hall *et al* 2009, Santana *et al* 2014) and are in line with the proposal of this paper.

The technical development of a sound automatic identification system for the *Anastrepha fraterculus* group was discussed in a preceding paper (Faria *et al* 2014). They evaluated image description approaches for encoding color and shape properties of images of wings and aculei for three *Anastrepha* species into feature vectors.

Some fruit flies (Tephritidae) are pests of quarantine importance, such as those in the genus *Anastrepha*. Despite extensive studies, the taxonomy of some *Anastrepha* groups is not fully resolved and poses many practical difficulties. Correct identification of species of economic importance in the *fraterculus* group is based mostly on female morphology. *Anastrepha* species identification is crucial for the implementation of management initiatives, control programs and quarantine restrictions. The purpose of this study is to apply image analysis and machine learning techniques as automated identifiers of three closely related *Anastrepha* species, namely, *Anastrepha fraterculus* (Wiedemann), *Anastrepha obliqua* (Macquart), and *Anastrepha sororcula* Zucchi, based on wing and aculeus images.

Material and Methods

Fruit fly samples

Anastrepha fraterculus, *A. obliqua*—both species are being considered in sensu lato—and *A. sororcula* adults were trapped in Monte Alegre do Sul (22°40′S, 46°40′W), Presidente Prudente (22°08′S, 51°23′W) and Monte Alto (21°15′S, 48°30′), state of São Paulo, Brazil. In addition, *A. fraterculus* were reared from guava, loquat, and peach, *A. obliqua* from guava and mango, and *A. sororcula* from guava. Specimens were identified by MFSF, based on the aculeus tip (Zucchi 2000). Individuals of each species with aculeus and right wing in good conditions were selected for image acquisition and analysis.

Image acquisition

The oviscape of each specimen was dissected and treated in a solution of 10% KOH for 12 h. The aculeus was then removed, placed ventral side up on a microscope slide with glycerin, and covered by a cover slide. After being imaged, the aculei were kept in polyethylene tubes with glycerin. The right wing of each specimen was removed, mounted on a slide with Euparal® and covered by a cover slide. The slides were photographed with a Nikon® DS-Fi1 camera. For the aculei, the camera was attached to a Nikon® microscope (10× and 40× objectives) (Fig 1); for the wings, we used a Nikon® SMZ 1500 stereomicroscope (1.5*x* objective) (Fig 2). Voucher specimens are deposited in the collection of the Department of Entomology and Acarology, ESALQ, Piracicaba, SP, Brazil.

Image analysis

The workflow to classify the images into different species consists of four steps: segmentation, description, training in classification and validation. The selected workflow is very similar for wings and aculei, only with a different descriptor.

To separate the structures from the background in the images, we used Otsu's segmentation method. This procedure consists of labeling all pixels in an image as either object or background. It finds an optimum threshold value to minimize the intra-class variance in levels of a gray image with the aim of separating object from background (Otsu 1979), producing a binary map that shows the separation between target objects and background.

The segmented images were then described by means of different descriptors for wings and aculei. For wings, the local color histogram (LCH) was used. This is an extension of the global color histogram (GCH) traditional color descriptor, which quantizes/splits the color space of an image into 64 uniform values (bins) per subimage from a 4×4 grid and computes the number of pixels belonging to each bin to compose the final feature vector with 1024 (4×4×64) dimensions (Swain & Ballard 1991). For the aculei, the edge orientation auto-correlogram (EOAC-288f) shape descriptor was used. This classifies the image edges according to their orientation and to the correlation between neighboring edges. The final feature vector is composed of values

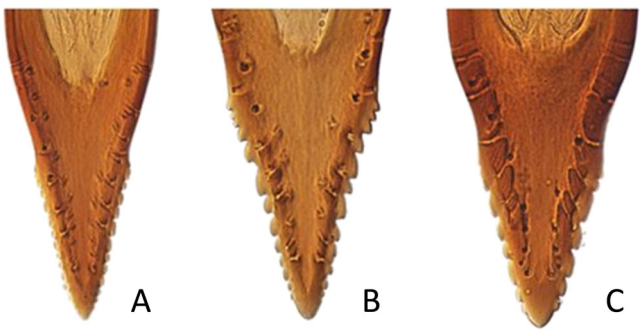

Fig 1 Aculei. *Anastrepha fraterculus* (**A**), *Anastrepha obliqua* (**B**), *Anastrepha sororcula* (**C**).

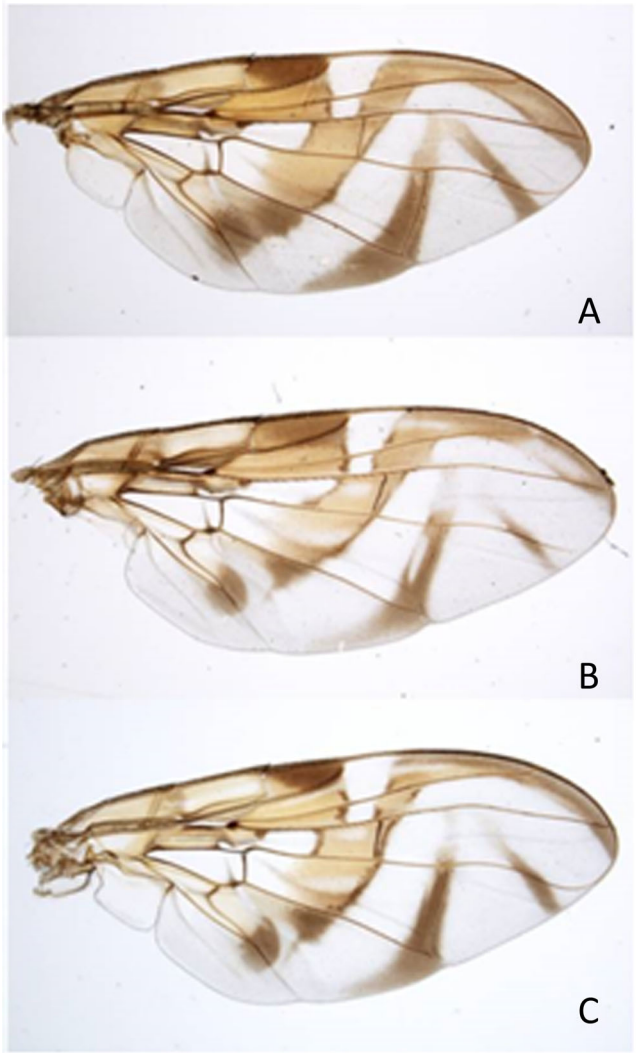

Fig 2 Wings. *Anastrepha fraterculus* (**A**), *Anastrepha obliqua* (**B**), *Anastrepha sororcula* (**C**).

extracted from this auto-correlogram. In this implementation, edge orientation quantization has 72 segments of 5° each and four distance values (1, 3, 5, and 7). The final feature vector is composed of 288 values (Mahmoudi *et al* 2003).

After description, images from identified individuals were used to train a machine learning algorithm that establishes criteria to classify images into different species. The Support Vector Machine (SVM) method (Arribas *et al* 2011, Breisman 1996) was used. This algorithm constructs an optimal hyperplane or set of hyperplanes that maximize the margin between classes in the image descriptor multidimensional space. The instances on borders between the classes are called support vectors. Prior to use, this classification algorithm must be validated. We used fivefold cross-validation to assess the accuracy of the method proposed here. In this protocol, the image dataset is divided into 5 subsets. Then, four subsets are used for training the classification algorithm, and the remaining one is used for testing. By doing this iteratively, 5 subsets are classified, and accuracy can be calculated for all images.

Results

A total of 596 images were tested, with an overall accuracy of 94.47% (wing and aculeus images), a highly promising result.

In the aculeus analyses, 296 tests were performed. The shape descriptor achieved an accuracy of 94.94 ± 3.63% (only 15 errors) (Table 1). Images of *A. fraterculus* had the highest number of correct identifications, with only one individual being erroneously identified as *A. obliqua* out of 98 tests. Images of *A. obliqua* had the highest number of errors (eight errors out of 99 tests); in every case, these were erroneously identified as *A. sororcula*.

The wing image analyses using a color descriptor also showed positive results, with a mean classification accuracy of 94% ± 2.88%. A total of 300 tests were performed, out of which only 17 resulted in species identification errors (Table 2). However, some images of all three species were misidentified. Higher identification rates were obtained for the *A. obliqua* and *A. sororcula* images (96% accuracy), whereas *A. fraterculus* images were identified with 91% accuracy.

Discussion

The identification of *Anastrepha* species is primarily based on morphological characters, mainly the aculeus tip. Examination of the aculeus tip requires dissection and subsequent analysis of its shape under a microscope, so that reliable identification to the species level requires taxonomists or trained entomologists. However, for some economically important species, identification is being based on molecular (Silva 2000, Smith-Caldas *et al* 2001) and morphometric (Araujo *et al* 1998, Bomfim *et al* 2011, Lopes *et al* 2013) analyses.

Image analysis is an alternative to facilitate identification of *Anastrepha* by non-taxonomists. However, no

Table 1 Identification of three *Anastrepha* species based on aculeus image analysis (SVM + EOAC—288f).

	A. fraterculus	*A. obliqua*	*A. sororcula*	Hits (%)	Total
A. fraterculus	97	1	0	98.97%	98
A. obliqua	0	91	8	91.91%	99
A. sororcula	0	6	93	93.93%	99

Rows show actual species and columns show species identities according to the classifier.

Table 2 Identification of three *Anastrepha* species based on forewing image analysis (SVM + LCH).

	A. fraterculus	*A. obliqua*	*A. sororcula*	Hits (%)	Total
A. fraterculus	91	6	3	91%	100
A. obliqua	2	96	2	96%	100
A. sororcula	1	3	94	96%	100

Rows show actual species and columns show species identities according to the classifier.

computerized image analysis technique for identifying *Anastrepha* species had been tested until Faria *et al* (2014). Among several classifiers tested by Faria *et al* (2014), whose best results were 78.3% for wing and 93.55% for aculeus images, we chose the SVM classifier to be used in this study. We obtained better species identification results using this classifier (88.8% based on wings and 96.0% based on aculeus images). Thus, using SVM classifier and higher number of images, percentage of correct identification was increased (consistently above 90%). Therefore, this technique holds promise to be very helpful to non-taxonomists who need to identify *Anastrepha* species by non-taxonomists.

In this study, the classification accuracy of image analysis was similar to that of morphometric analyses performed on exactly the same images. Multivariate morphometry of the aculeus and geometric morphometry of the wing showed 96.6 and 98.6% accuracy, respectively (Perre *et al* 2014), whereas the aculeus and wing image analyses achieved accuracies of 95 and 94%, respectively.

As stated above, aculeus tip shape is the main feature used to distinguish *Anastrepha* species; hence, we expected the aculeus image analyses to show better accuracy than the wing image analyses. However, the differences in accuracy of automatic identification based on these two structures were quite minor, indicating that wing and aculeus images can be considered similar in terms of efficacy. The few errors that did occur in the wing analyses may be due to high similarity in color patterns (indistinguishable to the human eye). Another possible reason for these errors is that the images were not taken on the same day, which may have produced subtle differences in lighting.

Generally, wing pattern is little used for species identification in the *Anastrepha fraterculus* group, as the differences in pattern and color are very elusive. Morphological identification of these species is therefore primarily based on microscopic examination of the aculeus tip. However, computerized image analysis of the wing effectively separated the three species using a color descriptor. This is an important finding, as preparing microscopic slides of wings for image capture is far easier than preparing aculei. Moreover, both females and males can be identified via wing images, whereas only female specimens can be identified based on aculei.

Identification of closely related species—such as those in the *Anastrepha fraterculus* group—is not a simple task, and so we expected that the first test of an automatic identification tool would require adjustments. The obtained results are promising for *Anastrepha* identification, as we were able to separate three closely related species with high classification accuracy. This is therefore the first step towards the development of an automatic identification tool.

Considering that only seven species of *Anastrepha* are economically important in Brazil (Zucchi 2000), the development of a tool that permits rapid identification of these species which can be used by non-professional taxonomists should be encouraged. In this sense, using the wing for identification instead of the aculeus is crucially important. In the case of quarantine pests, automating their identification could greatly expedite the work of quarantine agents, especially in ports and airports, highlighting the importance of this study for the identification of pest insects.

Acknowledgments PP was supported by a graduate scholarship from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). LRJ was supported by graduate and post-doctoral grants by Fapesp (grants 09/54806-0 and 14/16082-9). RAZ, RST, and AR are research fellows of the National Council for Scientific Research (CNPq). This study is part of the dissertation presented by PP to obtain a M.Sc. in Entomology at ESALQ/University of São Paulo.

References

Araujo EL, Nascimento FM, Zucchi RA (1998) Utilização de análise discriminante em estudos taxonômicos de moscas-das-frutas do gênero *Anastrepha* Schiner, 1868 (Diptera, Tephritidae). Sci Agric 55: 105–110

Arbuckle T, Schrder S, Steinhage V, Wittmann D (2001) Biodiversity informatics in action: identification and monitoring of bee species using ABIS. International Symposium for Environmental Protection 425–430

Arribas JI, Sánchez-Ferrero GV, Ruiz-Ruiz G, Gómez-Gil J (2011) Leaf classification in sunflower crops by computer vision and neural networks. Comput Electron Agric 78:9–18

Baylac M, Villemant C, Simbolotti G (2003) Combining geometric morphometrics with pattern recognition for the investigation of species complexes. Biol J Linn Soc 80:89–98

Bomfim ZV, Lima KM, Silva JG, Zucchi RA (2011) A morphometric and molecular study of *Anastrepha pickeli* Lima (Diptera: Tephritidae). Neotrop Entomol 40:587–594

Bortolus A. (2008) Error cascades in the biological science: the unwanted consequences of using bad taxanomy in ecology. AmBio 37:114–118

Breisman L (1996) Bagging predictors. Mach Learn 24:123–140

Faria FA, Perre P, Zucchi RA, Jorge LR, Lewinsohn TM, Rocha A, Torres RS (2014) Automatic identification of fruit flies (Diptera: Tephritidae). J Vis Commun Image Represent 25:1516–1527

Gaston KJ, O'Neil MAL (2004) Automated species identification: why not? Philos Trans R Soc B 359:655–667

Godfray HCJ (2002) Challenge for taxonomy: the discipline will have to reinvent itself if it is to survive and flourish. Nature 417:17–19

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11:10–18

Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. Syst Biol 54:852–859

Hebert PDN, Cywinska AS, Gall L, De Waard JR (2003) Biological identification through DNA Barcodes. Proc R Soc Lond B 270:313–332

Hortal J, De Bello F, Diniz-Filho JAF, Lewinsohn TM, Lobo JM, Ladle RJ (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. Annu Rev Ecol Evol Syst 46:523–549

LaSalle J, Wheeler Q, Jackway P, Winterton S, Hobern D, Lovell D (2009) Accelerating taxonomic discovery through automated extraction. Zootaxa 2217:43–55

Lopes GN, Arias OR, Cônsoli FL, Zucchi RA (2013) The identity of specimens of the *Anastrepha fraterculus* complex (Diptera, Tephritidae) with atypical aculeus tip. Neotrop Entomol 42:618–627

MacLeod N, Genfield M, Culverhouse P (2010) Time to automate identification. Nature 467:154–155

Mahmoudi F, Shanbehzadeh J, Eftekhari-Moghadam A, Soltanian-Zadeh H (2003) Image retrieval based on shape similarity by edge orientation autocorrelogram. Pattern Recognit 36:1725–1736

Otsu N (1979) A threshold selection method from gray-level histograms, systems. IEEE Trans Man Cybernet 9:62–66

Perre P, Jorge LR, Lewinsohn TM, Zucchi RA (2014) Morphometric differentiation of fruit fly pest species of the *Anastrepha fraterculus* group (Diptera: Tephritidae). Ann Entomol Soc Am 107:490–495

Russell K, Do M, Huv J, Platnick N (2007) Introducing SPIDA-web: wavelets, neural networks and Internet accessibility in an image-based automated identification system. Syst Assoc 74:131–152

Santana FS, Costa AHR, Truzzi FS, Silva FL, Santos SL, Francoy TM, Saraiva AM (2014) A reference process for automating bee species identification based on wing images and digital image processing. Ecol Inform 24:248–260

Silva OLR (2000) Controle do trânsito de hospedeiros de moscas-das-frutas. In: Malavasi A, Zucchi RA (eds) Moscas-das-frutas de importância econômica no Brasil: conhecimento básico e aplicado. Holos Editora, Ribeirão Preto, pp 29–40

Smith-Caldas MRB, McPheron BA, Silva JG, Zucchi RA (2001) Phylogenetic relationship among species of the *fraterculus* group (*Anastrepha*: Diptera: Tephritidae) inferred from DNA sequences of mitochondrial cytochrome oxidase I. Neotrop Entomol 30:565–573

Swain M, Ballard D (1991) Color indexing. Int J Comput Vis 1:11–32

Watson AT, O'Neill MA, Kitching IJ (2003) A qualitative study investigating automated identification of living macrolepidoptera using the Digital Automated Identification SYstem (DAISY). Syst Biodivers 1: 287–300

Weeks PJD, Gaston KJ (1997) Image analysis, neural networks, and the taxonomic impediment to biodiversity studies. Biodivers Conserv 6: 263–274

Weeks PJD, O'Neill MA, Gaston KJ, Gauld ID (1999) Automating insect identification: exploring the limitation of a prototype system. J Appl Entomol 123:1–8

Wheeler QD, Raven PH, Willson EO (2004) Taxonomy: impediment or expedient? Science 303:285

Zucchi RA (2000) Taxonomia. In: Malavasi A, Zucchi RA (eds) Moscas-das-frutas de importância econômica no Brasil: conhecimento básico e aplicado. Holos Editora, Ribeirão Preto, pp 13–24