Morphology-based cyber identification engine to identify ants of the Tetramorium caespitum/impurum complex (Hymenoptera: Formicidae)

Florian M. STEINER, Birgit C. SCHLICK-STEINER & Karl MODER

Abstract

We present a morphology-based identification engine for ants of the Tetramorium caespitum/impurum complex, embedded in the world wide web. Using classificatory discriminant analysis the facility allows one to discriminate workers of nine very similar species, namely T. caespitum (LINNAEUS, 1758), T. impurum (FOERSTER, 1850), T. hungaricum RÓSZLER, 1935, T. tsushimae EMERY, 1925, and T. spp. A - E. Identification requires 21 characters, captured with high-precision morphometry. The identification engine includes verbal and graphic character definitions as well as an entry mask for morphometric data. Upon data entry, probability values for the sample's membership to each of the nine species are returned. We recommend users to apply and put on record a probability cut-off.

Key words: Online identification engine, interactive species discrimination, world wide web, cryptic biodiversity, high-precision morphometry, classificatory discriminant analysis, Formicidae, ants, Tetramorium.

Introduction

Cryptic species, i.e., species that are morphologically nearly indistinguishable from others, pose most difficult challenges to determination. Molecular approaches can be helpful in this process. However, the identification of dried insect vouchers – crucial for research and society (SUAREZ & TSUTSUI 2004) – by molecular means is still in its infancy (e.g., HAJIBABAEI & al. 2005). Morphology will continue to hold a central position in routine insect identification.

Quantitative, metric data are indispensable for morphological-based taxonomy and determination of most similar species (SEIFERT 2002). The numbers of required differential characters rise with increasing intraspecific variation and decreasing interspecific differences. Frequently, multivariate statistical analyses are necessary for reliable discrimination (e.g., SEIFERT 2002, KLIMOV & al. 2004).

The idea of morphology-based digital determination is not new. Since three decades applications have been created (FERAL 1998). WHEELER (2004) lists "sophisticated interactive identification keys" as one of the seven immediate information technology opportunites taxonomists should make use of. Today a wide range of web-sites offers online versions of traditional dichotomous keys, partly clickable and thus transporting the user from couplet to couplet (ant examples: Australian subfamilies, SHATTUCK & BARNETT 2001; African Tetramorium, TAYLOR 2004; subgenus key of Polyrhachis, HASHIMOTO 2006). Other sites offer online determination by prompting the user to enter mostly qualitative information on a set of diagnostic characters and return species names that match the combination (e.g., European birds, SURFBIRDS.COM 2006; South Florida gastropods, UNIVERSITY OF IOWA 2005). As far as we know, however, there are only few sites for online determination by use of multivariate statistics procedures (e.g., bee mites, KLIMOV 2005).

We present a morphology-based identification engine, embedded in the world wide web, to identify workers of the Tetramorium caespitum/impurum complex, native to the Palearctic region. Recent multidisciplinary work (SCHLICK-STEINER & al. 2006) revealed that the complex encompasses at least nine species instead of four as formerly believed. These ants are among the most widespread and abundant ants in open habitats of the Palearctic region. They exert significant ecological influence by, among others, collecting seeds, entertaining trophobiotic relationships with sap suckers, feeding on carrion, and active hunting (SEIFERT 1996). At least two species of the T. caespitum/impurum complex have successfully established invasive populations in the Nearctic region (SCHLICK-STEINER & al. 2006, STEINER & al. 2006). Reliable species level determinations are needed, among others, for research of biogeography, faunistics, biodiversity, ecology, conservation biology, invasion biology and sociobiology. Extreme morphological similarity and considerable intraspecific variation of these species necessitates discrimination by high-precision morphometry and classificatory discriminant analysis.
Materials and methods
The Tetramorium caespitum/impurum complex

BOLTON (1976, 1977, 1979, 1980) outlined the Palearctic T. caespitum (LINNAEUS, 1758) group by means of morphological characters and allocated 55 species (BOLTON 1995). Within this group, the species with strongest resemblance to Tetramorium caespitum and T. impurum (FOERSTER, 1850) have been termed the T. caespitum/impurum complex (SCHLICK-STEINER & al. 2006). Ants of the T. caespitum/impurum complex have challenged discrimination and taxonomy for a long time (e.g., KUTTER 1977, SEIFERT 1996). Workers range from small to large and from light brown to black; the head is often strongly, less frequently weakly rugulose; the mesosoma bears longitudi- nual rugulae; propodeal spines are moderately short; the dorsal surfaces of petiole and postpetiole are finely sculpted or nearly smooth; and the first gastral tergite shows a centrally positioned measuring distance on vertex; measured in dorsal view across level of anterior eye margins; a centrally positioned measuring distance of 200 - 300 µm is divided by the number of rugae crossing it; the specific length of the measuring distance is adapted to individual size: in order to increase accuracy it should be as long as possible, but strongly convex surfaces should be avoided – i.e., it must not exceed the carinulae continuing the frontal carinae backwards. Carinae/carinulae just touching the measuring line and those exactly at its endpoints are counted as 0.5.

EL Maximum large diameter of elliptic eye. All structurally defined ommatidia, pigmented or not, are included.

EW Maximum small diameter of elliptic eye. All structurally defined ommatidia, pigmented or not, are included.

FL Maximum anterior distance of frontal carinae. If FL is not defined because frontal carinae converge frontal anterior the FRS level, FL is taken to be equal to FRS.

FRS Distance between frontal carinae immediately cau- dal to posterior intersection points between frontal carinae and lamellae dorsal to torulus. If these dor- sal lamellae do not laterally surpass the frontal carinae, the imaginary line between the deepest points of scape corner pits may be taken as the reference line. These pits take up the inner corner of the scape base when the scape is fully switched caudad and produce a dark triangular shadow in the lateral frontal lobes immediately posterior to the dorsal lamel- lae of the scape joint capsule. Only measured if FL is not defined.

ML Mesosoma length measured in dorsal view from caudalmost portion of propodeum to anterodorsal corner of pronotal slope (i.e., where coarse structure of dorsal pronotum and fine structure of ante- rioromost pronotum meet); equivalent measuring also possible in lateral view.

MPPL Distance between summit of caudodorsal corner of propodeal lobe and imaginary line from center of propodeal stigma to anterodorsal corner of ventral margin of metapleuron; calculated as $\cos \epsilon = (PLST^2 + MPST^2 - MPPL^2) / (2 * PLST * MPST)$ (Fig. 1C).

MPSP Maximum distance between anterodorsal corner of metapleuron and tip of spine (Fig. 1C).

MPST Maximum distance from center of propodeal stigma to anterodorsal corner of ventral margin of metapleuron (Fig. 1C).

MW Maximum mesosoma width.
Fig. 1: Morphometric characters for the optimal discrimination of nine species of the Tetramorium caespitum/impurum complex, on the petiole (A), the mesosoma (B) and the posterior part of the mesosoma (C). For verbal definitions see text.

PEH Maximum petiole height. The straight section of the ventral petiolar profile at node level is the reference line perpendicular to which the maximum height of the petiole node is measured. The dorsal measuring point is not necessarily positioned at the node top.

PENL Maximum length of petiolar node from dorsocaudal corner of petiole to anterior petiolar slope. The measuring line is directed orthogonal to a reference line formed by the dorsalmost linear part of the anterior petiolar slope (Fig. 1A).

PEW Maximum width of petiole.

PLST Distance between center of propodeal stigma and summit of caudodorsal corner of propodeal lobe (Fig. 1C).

PnHL Length of longest hair on corner of pronotum; arithmetic mean of both sides.

PoOc Postocular distance. Using the cross-scaled ocular micrometer the head is adjusted to the measuring position of CL; caudal measuring point: median posterior margin of head as average between peaks and valleys of microsculpture; frontal measuring point: median head at the level of the posterior eye margin; average of left and right postocular distance is calculated.

PosSPI Orthogonal distance of uppermost point of spine to a ventral reference line of mesosoma. The ventral reference line is brought to coincidence with the horizontal line of the cross-scale at magnifications of about 100×. Then, at a magnification of ≥ 250×, the mesosoma is carefully tilted to a position in which the ventralmost point of the metapleuron and the tip of the spine are at the same focal level. Note that this character is not necessarily measured in lateral view, but frequently in a dorsolateral view (Fig. 1B).

PosSPu Measured in same adjustment as for PosSPI and orthogonal to ventral reference line. The distance of the uppermost point of spine to the dorsalmost point of propodeum. Note that this point is in the given adjustment usually behind the median line of propodeum and can only be found after focusing movements. With the graduated scale of the ocular micrometer kept perpendicular in the visual field, focusing cannot induce a parallax error (Fig. 1B).

PPH Maximum height of postpetiole measured perpendicular to a line defined by the suture between dorsal and ventral sclerites.

PPW Maximum width of postpetiole.

SLd Maximum straight line scape length in dorsal view, excluding articular condyle; taken as arithmetic mean of both scapes. The scape is tilted to a position perpendicular to the swivelling plane of the funiculus segment.

SPWI Maximum distance between outer margins of spines; measured in dorsofrontal view.

The total classification error of the optimal discrimination of the primary data (i.e., the data used for finding the optimal discrimination as given in Materials and methods) is 0.152. As it can be concluded from Tab. 1, the classification error concerning single species varies from 0 for T. hungaricum to 0.328 for T. caespitum.

Bootstrapping (50 replicates, 100 calculations per replicate) of the calculation of a discriminant function for the optimal character combination on the whole data set minus one randomly selected worker, and subsequent classification of this single worker, results in an average error of 0.186 (minimum 0.140, maximum 0.230). Hence, the discrimination appears robust enough to produce similar results when applied to other than the primary data.

The cyber identification engine uses the selected sets of characters for discriminating the nine species via classificatory discriminant analysis (SAS INSTITUTE 2004). Calculations and identification are automated by Java programs, accessible as Digital supplementary material to this article, at the web pages of Myrmecologische Nachrichten / Myrmecological News. The engine's URL is <http://homepage.boku.ac.at/h505t3/DiscTet>. The identification engine includes verbal and graphic character definitions as well as an entry mask for morphometric data. Upon data entry into the entry mask (Fig. 2) and clicking "identify" the user is returned probability values for the sample's membership to each of the nine species, the highest value being highlighted. Data of one T. caespitum specimen can be viewed and used for exploring the engine by clicking the "example" button.
Correct use of the cyber identification engine

Two requisites assure reliable operation of the engine. (1) Correct pre-identification of a sample as belonging to the *T. caespitum*/*impurum* complex. For this purpose we recommend standard keys, e.g., SEIFERT (1996), RADCHENKO & al. (1998), where these samples key out as either *T. caespitum* or *T. impurum*. (2) Character definitions must be accurately implemented by use of the stereo microscope, at maximum magnification available. Consult SEIFERT (2002) for how to avoid common errors in high-precision morphometry.

A cut-off facilitates the interpretation of the probability values (\(P\)) for a sample's membership to the nine species. There is no objective criterion for defining a cut-off, but extrapolation from the primary data may help. Note however that extrapolation from the classification table (Tab. 1) just allows a rough estimate. Every cut-off discards the identifications of a certain portion of the sample (those with the highest \(P\) for any of the nine species being lower than the cut-off level). Thus a trade-off must be found between the cut-off of \(P\), the corresponding portion of unidentified samples, and the classification error for the identified samples. A low classification error requires a high cut-off of \(P\) and produces a high portion of discarded identifications. Figure 3 illustrates the trade-off for the primary data. If, for example, a classification error of only 0.01 is envisaged, the necessary cut-off of \(P > 0.94\) discards half of the identifications. If an error of 0.1 appeared tolerable, a cut-off of \(P > 0.65\) would discard only 20% of the identifications. Whatever cut-off level is chosen, the value should be put on record. When \(P\) exceeds the desired cut-off for none of the species, we recommend a cautious interpretation ("one of nine species of the *T. caespitum*/*impurum* complex"). Should one of the \(P\) values come close to the cut-off, the two species with a sum of \(P\) above the cut-off might be indicated (e.g., "*T. caespitum* or *T. sp. B*").

Although the primary data used for establishing the presented discrimination were collected from the widest geographic range possible, the identification engine might suffer from incomplete coverage of the morphological variation of the single species. This tentativeness, however, is inherent in any identification system. At present only tentative maps of the species ranges are available (SCHLICKSTEINER & al. 2006). In a future version of the cyber identification engine it may be possible to add the geographical origin of a sample to the set of discriminative characters and thus reduce the number of morphometric characters needed.
Fig. 3: Portion of identified and of correctly identified of all samples of the primary data (upper diagram), and portion of the correctly identified of the identified samples (lower diagram) dependent on the chosen cut-off of probability of samples' memberships of species, for single workers (n = 473). Broken lines connect portions of correctly identified of the identified samples of 0.90, 0.95, and 0.99 with corresponding probability cut-offs, and with corresponding portions of identified samples of all samples.

Another unsolved problem is the possible existence of further cryptic species. SCHLICK-STEINER & al. (2006) present some samples which may reflect pronounced intraspecific variation or constitute independent entities. The determinations suggested by the cyber identification engine could thus require a future split-up. Again, this uncertainty pertains to any taxonomic work and determination routine. For all these reasons, the user should strive to support worker identification by an analysis of male genitalia (see SCHLICK-STEINER & al. 2006) and / or by mitochondrial DNA sequences. As of July 2006, a total of 439 sequences, of 1113 bp of the gene cytochrome c oxidase subunit I, COI, and 454 bp of cytochrome c oxidase subunit II, COII, are available at GenBank (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotides).

The online facility of MODER (2004) for the discrimination of *T. tsushimae* and *T. sp. E* (the latter referred to as "T. cf. caespitum") is still valid, but prone to pre-identification errors. Although there are currently no indications, North American samples for which the site of MODER (2004) was constructed could represent other species than *T. tsushimae* and *T. sp. E*. Therefore it is advisable to use the present cyber identification engine.

Performances and restrictions of the identification engine

What the cyber identification engine does achieve is to unburden the user from all the statistical effort necessary for identifying species of the *T. caespitum*/*impurum* complex. But the engine is no panacea. The onus of pre-identification and of morphometric analysis remain the responsibility of the user.

Acknowledgements

Our thanks go to Boku, University of Natural Resources and Applied Life Sciences Vienna, to the Austrian Federal Ministry for Education, Science & Culture, and to the Austrian Science Fund, FWF, for funding; to Herbert Zettel, for bearing responsibility for independent reviewing; to Donat Agosti, Alfred Buschinger, Erhard Christian, Ross Crozier, Bálint Markó, Martin Pfeiffer, Bernhard Seifert, Christian Stauffer, Andy Suarez, James Trager, and Phil Ward, for constructive criticism on the manuscript and the cyber engine. The late Stefan Schödl inspired our work on *Tetramorium* by his vivid interest.

Zusammenfassung

senden Merkmale sind online zugänglich. Nach Eingabe der morphometrischen Daten in das Eingabefenster erhält der Benutzer Wahrscheinlichkeitswerte für die Zugehörigkeit der Probe zu jeder der neun Arten. Wir empfehlen, für die Interpretation der Ergebnisse einen Grenzwert für die Wahrscheinlichkeitswerte zu verwenden und zusammen mit der Artbestimmung festzuhalten.

References


