A Novel Tool to Better Understand the Diversity of Life

Professor Thomas Wilke

Scientia

A NOVEL TOOL TO BETTER UNDERSTAND THE DIVERSITY OF LIFE

Completing the inventory of the Earth's biodiversity is a huge challenge. Supported by the German Research Foundation (DFG), **Professor Thomas Wilke** and his research group at Justus Liebig University Giessen in Germany are addressing this challenge head on. The team is developing a novel open-access and semi-automated taxonomy platform that will encourage collaboration towards the common goal of documenting and describing as yet unnamed species.

An Incomplete Inventory of Earth's Biodiversity

Learning what species inhabit our planet is a fundamental quest for biology. Taxonomists are the stewards of knowledge about the world's living and fossil species – they discover, describe, and classify 'taxa' – groups of similar organisms. Despite the key role that taxonomy plays in the conservation and management of biodiversity, our taxonomic knowledge is far from complete. Even at the most conservative estimate, there are more unknown species than known ones.

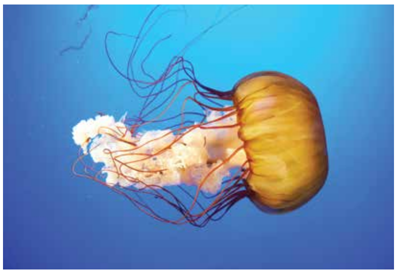
The science of taxonomy emerged in the 18th century with the Swedish botanist Carl Linnaeus. Linnaeus devised a two-part Latin-based naming system, whereby every organism has both a genus and a species name (*Homo sapiens*, for example). Linnaeus's biological classification system has been adapted by researchers for more than 250 years, and to date, around 1.5 million species have been validly described.

The actual number of species on Earth, however, remains uncertain – but estimates suggest that it is in the region of between 3 to over 100 million species. Therefore, there is a vast number of species that we know absolutely nothing about, and notably, the rate of species description per taxonomist has not increased in recent years.

Quality versus Speed

Significantly reducing the knowledge gap in the number of undescribed species is a weighty challenge. The speed at which species description can be documented by taxonomists is a factor, and with increasing extinction rates, this is especially important. There is an inherent conflict between the two main interests of taxonomy – the quality of description and delimitation (determining the boundaries between species) and the speed of description and delimitation. Each comes with its own set of specific issues.

There are several obstacles to high-quality species description. Examples include a lack of appropriate samples to make taxonomic comparisons, insufficient assessment of the range of character variation, neglecting to assess relevant characteristics, and disagreements on species concepts.



> **‘proSDS integrates various data sets, for example, anatomical, 3D-morphological, genetic, ecological, and biogeographical information, and uses supervised machine-learning approaches for dynamically delimitating species.’**

In addition, there are also problems associated with the speed of description. These can be related to data (accessibility of information, differences in data formats, and issues with standards of coding of species’ characteristics) and to inference (lack of standardised algorithms, problems with missing data, and conflicts among datasets).

Compounding matters, current species delimitations tend to be ‘black and white’. That is, assignments are classified as either ‘no species’ or ‘species’. In reality, however, there are areas of grey. If we assume that the process of speciation (the formation of new and distinct species) is an evolutionary process that lasts hundreds of thousands or even million years, such discrete assignments can bias biodiversity estimations.

In addition, the current number of taxonomists is insufficient to document the remaining species within a reasonable time period. Many believe that taxonomy is in a state of crisis and scientists are concerned that numerous species will disappear before being named.

DFG ‘Taxon-Omics’ Priority Programme

The DFG is tackling this taxonomic crisis by establishing its Priority Programme (SPP) 1991 ‘*Taxon-Omics: New Approaches for Discovering and Naming Biodiversity*’. The €5.5 million initiative recognises the importance of developing innovative methods for dissemination and preservation of knowledge for future discoverability and re-use.

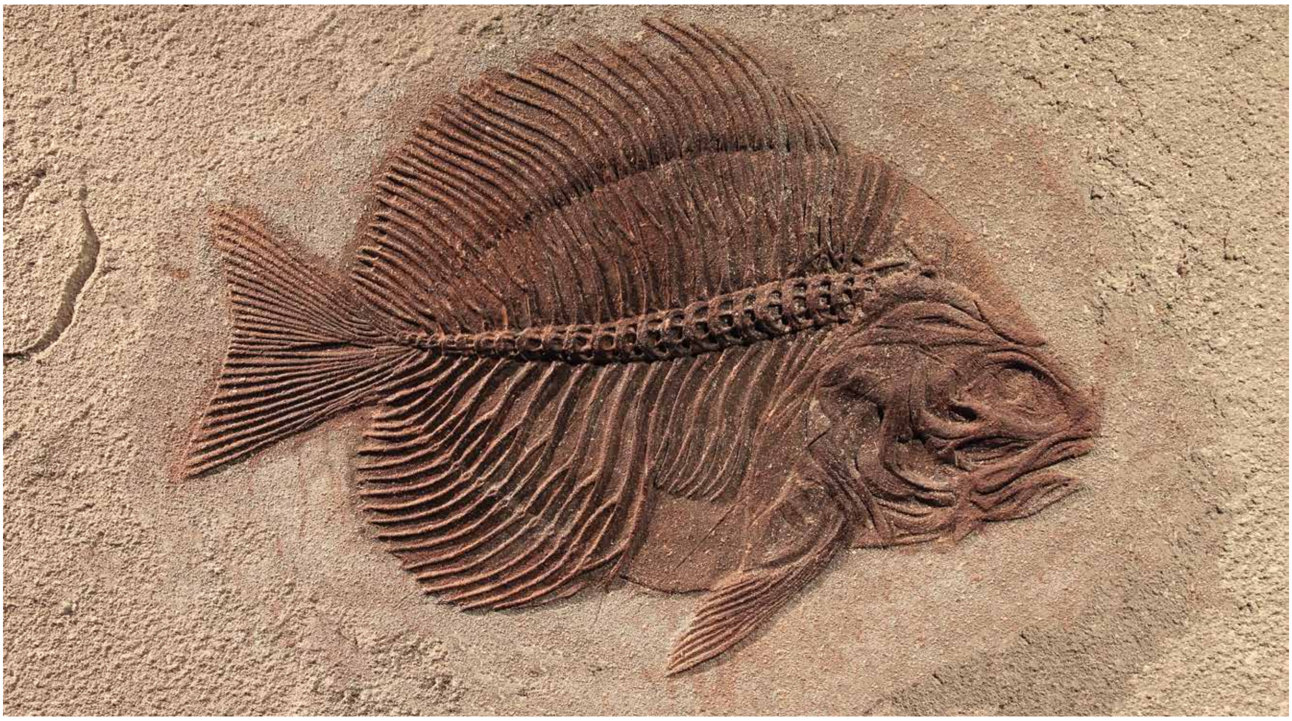
A promising solution to the taxonomic crisis is the integration of museum information with genetic data. DFG’s programme particularly supports new approaches that will increase the quality of species delimitation. It also aims to speed up the naming process, generate online identification tools, make novel use of natural history collections through ‘museomics’ (genomic data from historic specimens), and develop semi-automated analyses of specimens through machine learning.

A Novel Semi-Automated Species Discovery System

Professor Wilke and his team at the Institute of Animal Ecology and Systematics in Giessen hold one of 27 such research projects funded by the DFG’s Priority Programme. His team is developing a novel probabilistic and semi-automated species discovery system (or ‘proSDS’).

The ground-breaking tool will integrate non-genetic information with genetic data, utilising available information from museum collections and public databases. As Professor Wilke explains: ‘proSDS integrates various data sets, for example, anatomical, 3D-morphological, genetic, ecological, and biogeographical information, and uses supervised machine-learning approaches for dynamically delimitating species.’

Professor Wilke’s integrated approach was partly inspired by his postdoctoral experience at the Malacology Department of the Academy of Natural Sciences in Philadelphia, which boasts one of the largest mollusc collections worldwide. This work highlighted the huge value of these collections for making taxonomic decisions, but also the limitations of standard morphological approaches and the problem of missing genetic information. His experience led to a desire to better integrate museum materials with complex genetic (‘omics’) data sets.



The Snail Family Hydrobiidae

Professor Wilke and his team are drawing on a wealth of research expertise spanning 20 years into the taxonomy, systematics, and evolutionary biology of the microgastropod family Hydrobiidae. They are using this species-rich and taxonomically 'notoriously difficult' family as a model group with which to develop proSDS.

The tiny hydrobiid snails make up one of the largest gastropod families – with at least 900 valid species. Additionally, it's thought that the current number of named species may represent only one quarter of their actual diversity.

There are many gaps in the taxonomic knowledge of the hydrobiids. Genetic information is available only for a portion of species, and only a few large-scale comparative studies have been carried out. As Professor Wilke explains: 'The taxonomy of hydrobiids is clearly in need of revision in a comparative context using genetic, anatomical, ecological, and biogeographical information. Very likely, such integrated studies will further increase the number of described species.'

His team's first step will be to create a curated specimen-based reference database that includes a range of information on hydrobiid species (including genetic, biogeographic, ecological, anatomical, and 3D morphological data). As they will draw upon extensive previous research, only the 3D shell morphological scans will need to be generated from scratch. To obtain these scans, the team will take advantage of strong collaborations at European and North American museums, as well as the extensive collection of approximately 100,000 hydrobiid snails at the University of Giessen.

A Standardised and Scalable Tool

The proSDS team aims to address the problem of integrative species delimitation from a novel angle by utilising training datasets and a flexible machine-learning approach. Many of the system's features are innovative, including its ability to build on integrated data and to manage missing information. It can process mixed data types, retrieves information automatically from public databases and provides the user with probabilities that a specimen under investigation belongs to a known or novel species.

Initial evaluations of proSDS are very promising. The system demonstrates good performance, both for simulated and real data with high rates of correct species classifications over a wide range of variation within and between species.

An Open Source Resource

The team's aim is to assist scientists in making taxonomic decisions by estimating the probability that a specimen under investigation belongs to a known or novel species. The underlying machine-learning approach will also provide information for the individual contribution of the specific characteristics. Importantly, this may well alert taxonomists to potential conflicts among existing datasets.

In the future, proSDS will be applicable to a wide range of taxonomic groups. By providing open access to the tool, the team hopes that it will encourage collaboration towards the fundamental goal of furthering knowledge into the evolution, maintenance, and conservation of biological diversity.

Meet the researcher

Professor Thomas Wilke

Institute of Animal Ecology and Systematics

Justus Liebig University

Giessen

Germany

Professor Thomas Wilke graduated with a diploma in Biology with summa cum laude from Potsdam College and Humboldt University, Berlin, in 1989. He was awarded his PhD in Zoology from the University of Potsdam in 1994 and his habilitation in Zoology at Goethe University, Frankfurt am Main in 2001. Professor Wilke then held positions at the Academy of Natural Sciences, Philadelphia, and The George Washington University Medical Center, Washington DC, before taking a full professorship at Justus Liebig University Giessen in 2004. Here he is currently the Head of the Systematics and Biodiversity Group. Since 2005, he has also held the position of Vice President of the German Malacological Society (DMG).

CONTACT

E: tom.wilke@allzool.bio.uni-giessen.de

W: www.uni-giessen.de/wilke

FUNDING

German Research Foundation
US National Science Foundation
European Commission
German Academic Exchange Service
German Ministry of Higher Education and Research

FURTHER READING

T Wilke, M Haase, R Hershler, H-P Liu, B Misof, W Ponder, Pushing short DNA fragments to the limit: Phylogenetic relationships of 'hydrobioid' gastropods (Caenogastropoda: Rissooidea), Molecular Phylogenetics and Evolution, 2013, 66, 715–736.

T Hauffe, C Albrecht, T Wilke, Assembly processes of gastropod community change with horizontal and vertical zonation in ancient Lake Ohrid: a metacommunity speciation perspective. Biogeosciences, 2016, 13, 2901–2911.

D Delicado, T Hauffe, T Wilke, Ecological opportunity may facilitate diversification in Palearctic freshwater organisms: a case study on hydrobiid gastropods, BMC Evolutionary Biology, 2018, 18, 55.

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN