The Future of Systematics in Data-Centric Biology

Marine Biological Laboratory, October 25-28, 2017

Schedule Overview

Wednesday, Oct 25

Participants arrive in Woods Hole and check in at the front desk of Swope Hall. Dinner is available in Swope dining hall until 7:30 p.m. Woods Hole is a small town, so anyone arriving later in the evening (after about 9:30 p.m.) should make other plans for dinner.

Thursday, Oct 26

7:00	–	8:30	a.m.	Breakfast in Swope Dining
8:30	–	8:45	a.m.	Welcome and introduction
8:45	–	9:30	a.m.	Staffan Müller-Wille
9:30	–	10:15	a.m.	Betty Smocovitis
10:15	–	10:30	a.m.	Break
10:30	–	11:15	a.m.	Jen-Pan Huang
11:15	–	12:00	p.m.	Roberta Millstein
12:00	–	1:30	p.m.	Lunch
1:30	–	2:15	p.m.	David Remsen
2:15	–	3:00	p.m.	Nico Franz
3:00	–	3:30	p.m.	Break
3:30	–	4:15	p.m.	Beckett Sterner
4:15	–	5:15	p.m.	Discussion
5:15	–	7:30	p.m.	Dinner

Friday, Oct 27

7:00	–	9:00	a.m.	Breakfast
9:00	–	9:45	a.m.	Joeri Witteveen
9:45	–	10:30	a.m.	David Hibbett
10:30	–	10:45	a.m.	Break
10:45	–	11:30	a.m.	Ramon Rosselló-Móra
11:30	–	12:15	p.m.	Arvind Varsani
12:15	–	1:30	p.m.	Lunch
1:30	–	2:15	p.m.	Greg Morgan
2:15	–	3:00	p.m.	Robert Beiko
3:00	–	3:30	p.m.	Break
3:30	–	4:15	p.m.	Jens Kuhn
4:15	–	5:15	p.m.	Discussion
5:15	–	7:30	p.m.	Dinner

Saturday, Oct 28

Participants check out of Swope Hall by 10 a.m. and depart from Woods Hole.

Special thanks to the Marine Biological Laboratory
and the McDonnell Foundation for their generous support of this event.

Full Program

Thursday, October 26

8:30 – 8:45 a.m. — Welcome and Introduction

8:45 – 9:30 a.m. — Staffan Müller-Wille:

Linnaeus's Role in the History of Systematics

The Swedish botanist and physician Carl Linnaeus (1707–1778) is widely regarded as the founder of modern systematics. But what exactly was his contribution? I will argue that Linnaeus, at his time, shaped the future of systematics both with regard to its practice and theory. By introducing binomial nomenclature and a hierarchy of ranks, he created a system of "labels" and "containers" that allowed botanists and zoologists to collect and process information on plants and animals on an unprecedented scale. This achievement, which has often been addressed as "merely" pragmatic, had far reaching theoretical consequences also. It implied a demise of the idea of a scale of nature, and a strict distinction between "artificial" and "natural" systems. Systematics is an information science, I argue, yet changes in the way information about plants and animals is organized also imply changes in how we perceive nature.

9:30 – 10:15 a.m. — Betty Smocovitis:

Two Case Studies of Synthesis and Integration in the Origins of the "New" Systematics

This presentation uses two case studies from the history of plant systematics at a crucial moment in its history to get at the introduction of novel techniques, new problematics, interdisciplinary collaborations, as well as challenges. The two case studies center on the origins of "biosystematy," or the "new" systematics of the 1940s during the period of the evolutionary synthesis. The presentation is forward-looking in that it aims to shed light on current challenges and opportunities in plant systematics.

10:15 – 10:30 a.m. — Break

10:30 – 11:15 a.m. — Jen-Pan Huang:

Speciation continuum and the grey zone of species delimitation:
An empirical investigation using the Hercules beetle system

The speciation process can be a continuous process, where differences between evolutionary entities accumulate along different axes (e.g., genetic and phenotypic axes) with different rates. Therefore, species delimitation can be inconsistent across studies because different yard-sticks are chosen to mark species-level divergence. I show that speciation in the Hercules beetle system is a continuous process and that the species and subspecies designations are inconsistent. However, an integrative approach that incorporates different data types can help us more consistently delimiting species. I also test the consistency of species delimitation using molecular data under the multi-species coalescent model. Molecular species delimitation can be sensitive to the quality and quantity of the selected molecular markers. My finding unravels the need for a comprehensive review about what kind of evolutionary entities have been and can be delimited as species using solely molecular data.

11:15 – 12:00 p.m. — Roberta Millstein:

Populations and the Endangered Species Act: Lessons from the Grey Wolf

Since 1978, the U.S. Endangered Species Act has covered not only species and subspecies, but also "distinct population segments" (DPSs). And ever since then, the term DPS has been difficult to interpret and difficult to implement. US Fish and Wildlife (USFWS) Rules proposing the delisting of various grey wolf groupings (removing them from endangered species status) are a particularly egregious example of this. For grey wolves, an additional step of analysis was added, first analyzing the existence of a population and then performing a DPS analysis. However, there are concerns both with the concept of "population" used and the concept of DPS used for wolves. This talk traces the path of some of the relevant Rules and Proposed Rules and offers some suggestions for how population and DPS could be better characterized, including considering whether my own population and metapopulation concepts (Millstein 2010) could serve the purpose.

12:00 – 1:30 p.m. — Lunch

1:30 – 2:15 p.m. — David Remsen

The Use and Limits of Scientific Names in Biological Informatics

Scientific names serve as the primary, and often sole, means to link information about species to our understanding of biodiversity. This ubiquity implies a ciritical role in search, retrieval and integration of biological information in online information systems and, indeed, they form the primary means for locating species information. Names, and their underlying taxonomic definitions, however, are prone to instability and ambiguity. These features can profoundly and negatively impact the use of names as identifiers when employed as keywords to retrieve relevant information related to a taxon. Precision and recall are two measures of relevance that are directly impacted and they present different and distinct challenges to both users and providers of biodiversity data online. Addressing these challenges requires both technical and social engineering solutions.

2:15 – 3:00 p.m. — Nico Franz:

Machine-scalable solutions for future taxonomy
must begin with a philosophical recommitment to embrace trained judgment.

For nearly 10 years I have worked on the taxonomic concept approach – see https://doi.org/10.1093/sysbio/syw023. This approach offers a more granular syntax than the Linnaean system to align the concepts represented in multiple conflicting taxonomic or phylogenetic hierarchies, using the semantics of Region Connection Calculus articulations (congruent, includes, overlaps, etc.) to achieve integration. We have designed a powerful, custom logic reasoning tool that provides the desired integration services for a wide range of use cases. However, doing so requires experts to be sufficiently confident in making explicit empirical commitments regarding the identities of terminal and node concepts in the source hierarchies to be integrated. In other words, the challenges in promoting this approach run deeper than just learning to use a new tool. At root, they signal that systematists need to recommit ourselves to expressing the intentionality of systematic concepts, i.e., to embrace an integration culture where trained judgments are permitted to drive the integration process "top-down".

3:00 – 3:30 p.m. — Break

3:30 – 4:15 p.m. — Beckett Sterner
Biodiversity Informatics and the Epistemology of Taxonomic Concepts

The basic aim of biodiversity data aggregators, such as GBIF or iDigBio, has been to support data discovery. In this regard, they have been immensely successful at providing centralized portals to information about specimen vouchers and occurrence observations from organizations across the world. However, the names-based paradigm they currently embrace has fundamental limitations when we consider next-generation services that rely on high-quality data packages. In order to represent and reason over taxonomic disagreement and uncertainty, we need a more sophisticated way to formalize the meanings of taxonomic names. We also need new social mechanisms to catalyze the production of machine-readable representations of taxonomic knowledge. In this talk, I outline a solution by contrasting the epistemology of taxonomic concepts with the “realist” methodology behind the Open Biomedical Ontology program.

4:15 – 5:15 p.m. — Discussion

5:15 – 7:30 p.m. — Dinner

Friday, October 27

9:00 – 9:45 a.m. — Joeri Witteveen:

Linnaean naming philosophies and identity politics

Naming biological taxa is dealing in uncertainty. Changing hypotheses of taxon circumscriptions can raise questions about taxon identity. If an original concept of a taxon is later determined to have been a composite of several taxa, then which of those new concepts refers to the original taxon with smaller boundaries? Linnaeus perspicaciously anticipated such issues and described a procedure to address them. Yet his method for linking names to taxa led to new problems in the context of the increasingly and data-intensive taxonomic enterprise of the nineteenth century. It became apparent that names could start to 'drift' from their original designations and needed to be 'anchored' firmly to nature. In this talk, I will give a historically-informed philosophical analysis of this transition in taxonomic reference systems. I focus on the problem of transitioning existing names from one reference system to the other, which became a cause of vehement debate around 1900. A close examination of this debate reveals how the fundaments of the contemporary codes of taxonomic nomenclature were shaped through an intricate mix of philosophical argumentation and sociopolitical dispute.

9:45 – 10:30 a.m. — David Hibbett:

Why isn't this the golden age of fungal systematics?

This should be the golden age of fungal systematics. Whole genome sequences of diverse fungal taxa are being produced at an unprecedented rate, and massive volumes of environmental sequence data are being generated by molecular ecologists. Thanks to these two data streams, as well as continuing work in traditional (PCR-based) fungal phylogenetics, both the deep branches and tips of the fungal tree of life are being resolved with greater confidence and detail than ever before. However, our rich new understanding of fungal phylogeny is not being efficiently translated into systems of taxonomic names. Thus, consumers of taxonomic information, including most biologists, educators, and members of the general public, are cut off from recent advances in fungal phylogenetics. I argue that there are two structural aspects of fungal systematics that are hindering the translation of trees into taxonomy: (1) use of Linnaean ranks, which complicates and inhibits the naming of clades, and (2) the requirement for physical type specimens for new species, which prevents naming of species based on environmental sequences. Recent initiatives in phylogenetic taxonomy and fungal nomenclature may yield mechanisms to overcome these barriers, but it remains to be seen if rank-and-file fungal taxonomists will accept unranked taxa or species whose existence is inferred only from sequence data.

10:30 – 10:45 a.m. — Break

10:45 – 11:30 a.m. — Ramon Rosselló-Móra:

Prokaryotic species and databases

Archaea and Bacteria have been traditionally classified using pure cultures grown in the laboratory as the only way to retrieve essential genomic and phenotypic information. The circumscription of species has been improved in parallel to the methodological developments in molecular biology, but conspicuously genotyping has improved much extensively than phenotyping. Species are well circumscribed by means of in silico genome comparisons as well as housekeeping gene phylogenies, all based on data that can easily be stored in databases. Phenotypically, taxonomy still requires important developments as the retrieved information still relies on old-fashioned tests not generating database cumulative data. Important developments in phenotyping are still necessary. However, Prokaryote taxonomy suffers from two major problems. The first is that most of the descriptions are based on single isolates, not offering a clear idea of the intraspecific metabolic and genetic diversity. A problem that we overcome with large-scale culturing linked to mass spectrometry. But the second problem is the lack of place for the uncultured and vast majority of prokaryotes. The new sequencing technologies allow, by metagenomic approaches, the detection of single population's genomes, representing genomic species susceptible of being classified even with higher standards than the common single strain species descriptions based on cultures and poor phenotyping. The future of prokaryotic taxonomy, under my view, relies essentially on the genomic definition of the units and their phenotypic inference, with an ulterior experimental testing; as well as with a digitalized form of the protologues.

11:30 – 12:15 p.m. — Arvind Varsani:

Viral taxonomy in a data rich world

Viruses are ubiquitous in nature and the most abundant entity on the planet. They infect all organisms in the three domains of life. Approximately 1% of viral sequence space has be catalogued with a heavy bias for viruses associated with diseases in plants and animals. Viruses can have either DNA or RNA genomes (singles-stranded or double-stranded) and do not have universal genes. Hence, the lack of universal genes and poorly populated databases makes viral classification a major challenge and viral taxonomy, to some extent, is lagging behind in a data rich would fueled by high throughput sequencing.

12:15 – 1:30 p.m. — Lunch

1:30 – 2:15 p.m. — Greg Morgan:

The Challenges of Prolific Horizontal Gene Transfer for Classification

I will discuss how long evolutionary histories of horizontal gene transfer (HGT) raise problems for classification using viral classification as an example. Prolific HGT leads to highly mosaic viral genomes that are neither neatly clustered into species, nor neatly classified by hierarchical classification schemes. We can minimize but not entirely eliminate these problems by (1) examining evolving functional units smaller than genomes and by (2) giving up the deeply entrenched requirement that every virus is a member of one and only one species. Hierarchical classification makes more sense when evolutionary processes create branching non-reticulated histories. Non-hierarchical classification makes more sense when there is significant reticulation and a branching history oversimplifies evolutionary history. Relatedly, comparing a new viral genome against a database of genomes could be more useful and predictive than classifying it as a member of a given viral species.

2:15 – 3:00 p.m. — Robert Beiko:

Microbial Taxonomy: Abandon All Hope?

Classifying microorganisms has historically been a Sisyphean task where new types of information capture the imagination of researchers, revolutionize (augment, confound) existing taxonomic systems, and eventually fall apart due to their intrinsic limits of resolution or other flaws. Genetic and genomic information arguably represent the last great hope of having a taxonomy that is consistent with evolutionary history and has some predictive value.

Is molecular information up to the task? The millions of nucleotides contained within microbial genomes do offer a detailed view of phylogenetic similarities among organisms, and can be exploited to produce comprehensive functional predictions for any microorganism. However, the tendency for genomes to participate in larger genetic transfer events, and variation in function at even the smallest degrees of divergence suggest that a hierarchical Linnean system of classification can never be adequate to the task. So what should we use instead?

3:00 – 3:30 p.m. — Break

3:30 – 4:15 p.m. — Jens Kuhn:

Virus taxonomy — trailing other taxonomies or avantgarde?

Prior to the advent of next-generation/metagenomic sequencing, virus taxonomy was an ultimately cozy subspecialty because of the very low number (3,000-4,000) of known viruses and the resulting even fewer taxa needed for their classification. Virus taxonomy therefore trails other biological taxonomies in terms of scope. However, recent technical advances resulted in the description of >1,400 novel and highly diverse viruses in a single publication, and several similar manuscripts are in production. Thus, rather than using "ancient" Codes and principles, virus taxonomy may have the advantage over other biological taxonomies of being forced to develop prospective and modern mechanisms for both classification and taxon naming. Here, I will summarize the status quo of virus taxonomy from an operational point of view."

4:15 – 5:15 p.m. — Discussion

5:15 – 7:30 p.m. — Dinner