Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D18G89

Title: Developments of the integrative hypothesis of specialization

Creator: Marco Mello - **ORCID:** 0000-0002-9098-9427

Affiliation: Universidade de São Paulo (www5.usp.br)

Contributor: Carsten Dormann, Sharlene Santana, Renata Muylaert, Cristina Kita, Nastaran Lotfi, Tanja Maria Straka, Francisco Aparecido Rodrigues, Pierre-Michel Forget

Funder: São Paulo Research Foundation (fapesp.br)

Funding opportunity number: 2023/03083-6

Grant: 2023/03083-6

Template: Digital Curation Centre (português)

Project abstract:

No Laboratório de Síntese Ecológica (SintECO), estamos comprometidos em estudar interações entre organismos de diferentes espécies. Com uma abordagem baseada na ciência de redes, em nosso projeto anterior financiado pela FAPESP e outros órgãos, desenvolvemos um modelo cognitivo inovador, a hipótese integradora da especialização (IHS), que joga luz sobre as regras de montagem de redes de interações. Originalmente concebido como um modelo gráfico, após uma transformação em modelo algorítmico que resultou em uma prova de conceito bem-sucedida, a IHS mostrou-se capaz de explicar como surgem as quatro principais topologias observadas em redes de interações na natureza. Agora, em uma nova fase do projeto, pretendemos consolidar a IHS e explorar seus desdobramentos. Primeiro, apresentaremos a IHS como um modelo discursivo, ou seja, uma teoria eficiente no sentido estrito, enquadrada na visão epistemológica semântica. Segundo, investigaremos aplicações práticas da IHS, especialmente relacionadas a interações socioecológicas que envolvem também uma dimensão humana. Assim, consolidaremos uma nova teoria que poderá ser usada como uma ferramenta não apenas para o estudo de interações interespecíficas, mas também de serviços ecossistêmicos. Com esses próximos passos, estamos nos comprometendo a avançar em nossa linha de pesquisa sobre as regras de montagem de sistemas ecológicos complexos, que já está começando a ter um impacto significativo em diversos campos de estudo. Nosso projeto tem ainda o potencial de ajudar a alcançar alguns dos Objetivos de Desenvolvimento Sustentável propostos pela Organização das Nações Unidas, seguindo também a perspectiva da Saúde Única.

Start date: 08-01-2024

End date: 08-01-2026

Last modified: 04-04-2024

Developments of the integrative hypothesis of specialization

Serão compilados e curados dados sobre interações ecológicas entre organismos de espécies diferentes (também conhecidas como interações interespecíficas), tanto positivas quanto negativas, abrangendo diversos táxons. Os principais tipos de interações em que focaremos são a polinização, a dispersão de sementes, a herbivoria, o parasitismo e as zoonoses. Também incluiremos dados sobre interações entre humanos, animais, plantas e microorganismos por uma perspectiva socioecológica. Por fim, geraremos dados *in silico* para simulações computacionais. Nossas principais fontes de dados empíricos são o banco de de dados do SintECO, composto por dados oriundos dos estudos feitos pela nossa equipe, dados compartilhados conosco por colaboradores e dados de acesso aberto disponibilizados em *data papers* e bancos de dados online.

Banco de dados

No SintECO, armazenamos um grande volume de dados, que foram coletados por nós mesmos, cedidos por colaboradores, minerados da literatura ou extraídos de repositórios abertos. São englobados mutualismos, antagonismos, comensalismos e relações tróficas. Nosso banco de dados está organizado na forma de quadros de dados (*data frames*) verticalizados em formato CSV ou TXT, produzidos de acordo com a filosofia *tidy data* (https://vita.had.co.nz/papers/tidy-data.html) e conectados entre si por chaves lógicas padronizadas que permitem o cruzamento de informações. Partindo desses cruzamentos, disponibilizamos dados brutos, dados processados, códigos para análise e tutoriais (*notebooks*) como suplementos dos artigos que publicamos. Boa parte desse material já está disponível na nossa conta principal do GitHub (https://github.com/marmello77).

Estes são os principais conjuntos de dados (*data sets*) contidos no nosso banco de dados, que continuaremos usando para testar empiricamente previsões deduzidas da IHS:

1. **NeoBat Interactions**. É o principal conjunto de dados compilado e curado pela equipe do SintECO, contendo interações de frugivoria e nectarivoria entre morcegos e plantas em toda a Região Neotropical. Além de contar com informações em uma grande variedade de escalas espaciais, temporais e filogenéticas, e ser organizado à luz da filosofia *tidy data*, ele é focado em morcegos (Mammalia: Chiroptera), táxon em que o pesquisador responsável e diversos dos pesquisadores associados são peritos. Temos compilado e usado essas informações de diferentes maneiras em vários artigos, além de monografias, dissertações e teses. Atualmente, ele conta com 2.571 registros de interações entre 93 espécies de morcegos e 501 espécies de plantas. Foi publicado na forma de *data paper* (https://doi.org/10.1002/ecy.3640).

2. **BatFly Interactions**. Esse é o mais novo conjunto de dados compilado e curado pela equipe do SintECO, contendo informações sobre interações entre morcegos e moscas ectoparasitas na Região Neotropical e seguindo a filosofia *tidy data* usada no NeoBat. Atualmente, conta com 3.518 registros de interações entre 281 espécies de morcegos e 307 espécies de moscas ectoparasitas das famílias Streblidae e Nycteribiidae. Foi publicado na forma de *data paper* (https://doi.org/10.1002/ecy.4249).

3. **Krasnov Database**. Conjunto de dados sobre interações entre mamíferos e pulgas na Região Neártica. Desde 2016, o Prof. Boris Krasnov (https://cris.bgu.ac.il/en/persons/boris-krasnov), da Ben-Gurion University of the Negev (Israel), deu-nos acesso a ele. Assim como o NeoBat e o BatFly, ele pode ser recompilado em diferentes escalas espaciais, temporais e ecológicas, o que lhe confere grande plasticidade de uso. Atualmente, conta com 1.200 registros de interações entre 102 espécies de mamíferos e 161 espécies de pulgas.

4. **Forget Database**. Conjunto de dados sobre interações de frugivoria e dispersão de sementes entre diferentes grupos de vertebrados e diferentes grupos de plantas em dezenas de localidades nas regiões Neotropical, Afrotropical e Indomalaia. Desde 2012, o Prof. Pierre-Michel Forget (https://mecadev.cnrs.fr/index.php?

post/Forget-Pierre-Michel), do Muséum national d'Histoire Naturelle (França), deu-nos acesso a ele. Assim como o NeoBat e o BatFly, ele pode ser recompilado em diferentes escalas espaciais, temporais e ecológicas, o que lhe confere grande plasticidade de uso. Atualmente, conta com 19.478 registros de interações entre 558 espécies de animais e 2.196 espécies de plantas.

5. **Big Bat Database**. Base de dados sobre biologia de morcegos, disponibilizada online apenas para participantes da rede de pesquisa Global Union of Bat Diversity Networks (https://www.gbatnet.org). Contém, dentre outras, informações sobre morfologia, traços funcionais, características da história de vida, bioacústica, comportamento, genômica, requisitos de habitat, parasitas, patógenos e tamanhos populacionais de morcegos no mundo todo. Também conta com dados sobre interações entre morcegos e humanos por uma perspectiva socioecológica.

6. **BeeFlow** (nome provisório). Base de dados curada pela equipe do SintECO e ainda não publicada, contendo diversas informações sobre polinização de lavouras ao redor do mundo, organizada à luz da filosofia *tidy data*. Os dados foram minerados a partir da literatura, através de revisões sistemáticas feitas de acordo com o protocolo PRISMA (http://www.prisma-statement.org), especialmente em sua versão EcoEvo. Foram incluídos dados sobre o tipo de lavoura, regime de manejo, geolocalização, contexto da paisagem, uso de agrotóxicos, plantações florais (*hedgerows* e *flower strips*), dentre muitos outros. Em sua versão atual, conta com dados extraídos de cerca de 500 artigos científicos.

7. **COMBINE**. Conjunto de dados de acesso aberto publicado como *data paper* (https://doi.org/10.1002/ecy.3344). Contém informações sobre 54 características de 6.234 espécies de mamíferos existentes e recentemente extintas no mundo todo, incluindo informações sobre morfologia, reprodução, dieta, biogeografia, hábito de vida, fenologia, comportamento, área de vida e densidade.

Análises de redes

Neste projeto, usamos a terminologia do livro Network Science (http://networksciencebook.com). Definimos como **rede** um sistema complexo formado por espécies de **recursos** e espécies de **consumidores** que exploram esses recursos. Generalizamos como consumo diferentes tipos de interações ecológicas positivas, negativas e neutras, que incluem, por exemplo, polinização, dispersão de sementes, herbivoria, parasitismo e zoonoses. Cada espécie é considerada como um **nó** da rede e cada conexão entre espécies é chamada de **elo**. Quando uma rede contém apenas um tipo de elo, ela é denominada **monocamada**; caso possa ter dois ou mais tipos de elos, ela se torna **multicamada**.

Seguindo a terminologia da IHS, cada evento de consumo de uma espécie de recurso por uma espécie de consumidor é chamado de **interação**, sendo que o peso da interação entre uma espécie de consumidor j e uma espécie de recurso i é dada pela força da associação entre elas. Essa força pode ser operacionalizada de diferentes formas, de acordo com o tipo de interação e os táxons usados como modelo. Sempre que há informação suficiente disponível, quando trabalhamos com redes na **escala local**, consideramos os pesos das interações e, portanto, analisamos **redes ponderadas**. Quando trabalhamos em escalas espaciais maiores, como um **bioma** ou **continente**, trabalhamos com **redes binárias**, que apresentam apenas informações do tipo 1/0 (conexão presente/ausente). Temos trabalhado com redes que variam muito em **tamanho** (número de nós), **grau** (número de elos) e **conectividade** (distribuição dos elos entre os nós). As redes na escala local com as quais lidamos costumam ter algumas dezenas de nós. Já as redes na escala do bioma ou da região biogeográfica atingem centenas ou milhares de nós.

Já que a IHS trata basicamente da estrutura de redes de interações, nosso foco serão as análises estruturais focadas

na **topologia**, considerando os arquétipos conhecidos como composta, aninhada, modular e gradiente. Usaremos as novas análises que desenvolvemos para estudar a topologia de redes empíricas e sintéticas, considerando **modelos nulos** criados a partir de ajustes dos parâmetros da IHS. Também usaremos **modelos gerativos** para estruturação de redes complexas a fim de produzirmos análises baseadas em modelos gerais que vão além de Erdõs-Rényi, como por exemplo Barabási-Albert e Watts–Strogatz, ou modelos nulos populares na Ecologia, como vaznull e Patefield.

Por fim, também faremos análises de redes no nível do nó, usando uma combinação de métricas de **centralidade** que, juntas, podem nos ajudar a avaliar a importância de cada elemento para a estrutura do sistema ao qual ele pertence. Pretendemos usar principalmente as seguintes métricas: grau relativo, proximidade, intermédio, autovetor, *pagerank*, acessibilidade, grau dentro do módulo e coeficiente de participação. Cada uma dessas métricas captura um aspecto diferente da conectividade e importância relativa do nó e pode ser usada como variável operacional para traduzir diferentes conceitos ecológicos.

A grande maioria dessas análises de redes será feita usando a **linguagem de programação R**, especialmente com base nos pacotes *bipartite*, *igraph, EMLN*, *mully*, *rMultiNet, muxViz* e *multinet*, além de novos pacotes e funções personalizadas (UDFs) desenvolvidas pela equipe do projeto. Também utilizaremos outras linguagens, como **Python**, **HTML** e **Markdown**, em casos específicos. Temos disponibilizado dados, códigos e notebooks em R e outras linguagens nas contas de **GitHub** e **Zenodo** dos membros do laboratório (https://marcomellolab.wordpress.com/software/). Muitas das nossas soluções, desenvolvidas em fases anteriores do projeto, já foram incorporadas ao pacote *bipartite* pelo Prof. Dormann, autor do pacote e membro da equipe. Pretendemos otimizar os códigos já produzidos em projetos anteriores, assim como os códigos que serão escritos nesta nova fase.

Os dados serão contextualizados através de metadados sobre (i) a natureza, a frequência e o contexto dos eventos de interação registrados na natureza ou no laboratório; (ii) os organismos envolvidos nas interações, sua taxonomia, história natural e traços funcionais; (iii) os locais onde essas interações foram registradas e suas condições ambientais (naturais, urbanas ou rurais); e (iv) as fontes de onde as interações foram compiladas. Junto com dados e metadados, disponibilizaremos documentos explicativos para potenciais usuários, ao estilo dos metadados usados nos *data papers* da revista Ecology.

Como o projeto será baseado em dados secundários e dados gerados *in silico*, além de não envolver pesquisa com seres humanos, não há questões éticas envolvidas, como por exemplo necessidade de pedir licenças de pesquisa a órgãos ambientais ou sanitários.

Os dados compilados pelo projeto serão posteriormente disponibilizados para o público na forma de *data papers* ou de conjuntos de dados depositados em repositórios de acesso aberto, em sua maioria usando-se licenças Creative Commons ou licenças próprias de cada revista ou repositório. Os códigos de programação gerados pelo projeto serão depositados e disponibilizados no GitHub sob licenças Creative Commons. Funções personalizadas (UDFs) e pacotes de R e Python serão eventualmente incorporados a pacotes maiores, como o *bipartite* para R, mantidos pelos nossos colaboradores.

Os dados e metadados serão armazenados na forma de *data frames* (planilhas verticalizadas), seguindo a filosofia *tidy data* (https://vita.had.co.nz/papers/tidy-data.html), em arquivos eletrônicos em formato TXT ou CSV. O backup será feito em servidores na nuvem via Google Drive institucional e também de forma física em SSDs externos plugados aos computadores do laboratório, gerenciados via software de backup Time Machine.

O acesso aos dados, durante a execução do projeto, será exclusivo aos membros do projeto. Depois que o projeto for encerrado e as nossas principais descobertas forem publicadas, os dados serão disponibilizados ao público

seguindo os padrões da filosofia *open science*, como comentado anteriormente.

Todos os dados sobre interações interespecíficas, assim como os metadados que os contextualizam, são de valor a longo prazo e, portanto, serão preservados tanto no laboratório quanto na forma de publicações de acesso aberto.

O plano de preservação a longo prazo está baseado nos backups feitos no laboratório (nuvem e físicos) e também na publicação dos dados na forma de *data papers,* assim como em seu depósito em repositórios estáveis e de acesso aberto.

Na forma de *data papers* e conjuntos de dados depositados em repositórios estáveis e de acesso aberto, com prioridade para plataformas como GitHub e Zenodo.

Restrições de acesso e uso serão aplicadas apenas até que sejam publicadas as principais descobertas do projeto.

Os responsáveis pelo gerenciamento de dados serão o pesquisador responsável pelo projeto e os demais pesquisadores associados.

Computadores de mesa Apple de alta performance, SSDs externos com conexão USB-C para backups físicos locais, e acesso institucional com espaço ilimitado ao Google Drive.
