Bayesian nonparametric prediction of the taxonomic affiliation of DNA sequences

Tommaso Rigon

UCLA - 2023-04-26

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

Co-authors...

Joint work with:

David Dunson
(Duke University)

Alessandro Zito
(Duke University)

...and collaborators

LIFEPLAN

A PLANETARY INVENTORY OF LIFE

- LIFEPLAN is a worldwide sampling program that aims to establish the current state of global **biodiversity**.
- In the Lifeplan team there are **statisticians**, **ecologists**, and **computer scientists**.
- We interact on a weekly basis to get feedback from the other "communities".

- PIs are: prof. Otso Ovaskainen, Tomas Roslin, David Dunson.

Species sampling modes

- This first part is a **gentle introduction** to BNP methods for species discovery, in an idealized and **simplified** setting.

- Let $V_1, \dots, V_n$ be some collection of **species** with frequencies $n_1, \dots, n_K$, sometimes called **abundances** in ecology.

- Suppose V_i are conditionally iid samples from a **species sampling model**, so that

$$(V_i \mid \tilde{p}) \overset{\text{iid}}{\sim} \tilde{p}, \qquad \tilde{p} = \sum_{h \geq 1} \pi_h \delta_{\theta_h},$$

  where $(\pi_h)_{h\geq 1}$ is a set of random probabilities and θ_h represent distinct species.

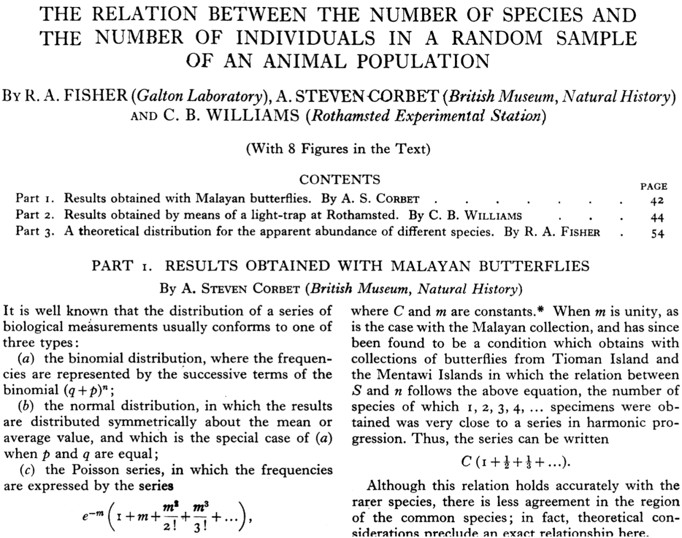
- The weights $(\pi_h)_{h\geq 1}$ are the **species proportions**.

- The discreteness of $\tilde{p}$ that identify $K_n = k$ distinct taxa, named $V_1^*, \dots, V_k^*$, with frequencies $n_1, \dots, n_k$.

Goals of this simplified analysis

- This **simplified** setting has a rich statistical history.
- Indeed, there are several quantities of ecological interest that one can try to obtain using abundances, for example:
- The **sample coverage**, namely the sum of the proportions of species that has been observed (Good 1953);
- The estimation and extrapolation of **accumulation curves** (aka rarefaction);
- The prediction of number of **unseen species** not observed in the current sample that we may observe in the future (Good & Toulmin, 1956);
- The estimation of **bio-diversity**, i.e. using the Simpson index, Shannon entropy, etc.
- and many more.

Fisher, Corbet and Williams (1943)



THE RELATION BETWEEN THE NUMBER OF SPECIES AND THE NUMBER OF INDIVIDUALS IN A RANDOM SAMPLE OF AN ANIMAL POPULATION

By R. A. FISHER (*Galton Laboratory*), A. STEVEN CORBET (*British Museum, Natural History*) AND C. B. WILLIAMS (*Rothamsted Experimental Station*)

(With 8 Figures in the Text)

CONTENTS

PART 1. RESULTS OBTAINED WITH MALAYAN BUTTERFLIES

By A. STEVEN CORBET (*British Museum, Natural History*)

It is well known that the distribution of a series of biological measurements usually conforms to one of three types:

(*a*) the binomial distribution, where the frequencies are represented by the successive terms of the binomial $(q+p)^n$;

(*b*) the normal distribution, in which the results are distributed symmetrically about the mean or average value, and which is the special case of (*a*) when p and q are equal;

(*c*) the Poisson series, in which the frequencies are expressed by the series

$$e^{-m}\left(1+m+\frac{m^2}{2!}+\frac{m^3}{3!}+\dots\right),$$

where C and m are constants.* When m is unity, as is the case with the Malayan collection, and has since been found to be a condition which obtains with collections of butterflies from Tioman Island and the Mentawi Islands in which the relation between S and n follows the above equation, the number of species of which 1, 2, 3, 4, ... specimens were obtained was very close to a series in harmonic progression. Thus, the series can be written

$$C\left(1+\tfrac{1}{2}+\tfrac{1}{3}+\dots\right).$$

Although this relation holds accurately with the rarer species, there is less agreement in the region of the common species; in fact, theoretical considerations preclude an exact relationship here.

Bayesian nonparametric priors

- The sampling distribution $\tilde{p}$ encodes all the **relevant information** but it is unknown, so we are interested in learning it from the data $V_1, \dots, V_n$.

- In the Bayesian framework, this amounts to the choice a **discrete nonparametric prior** for the sampling distribution $\tilde{p}$.

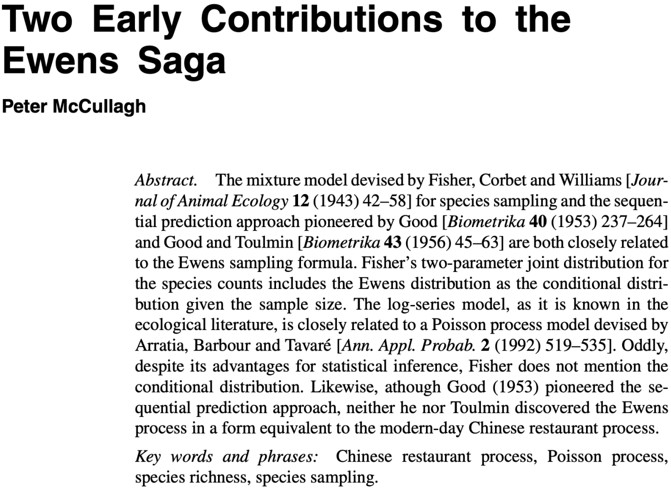
- Then, one can study the following posterior law

$$\tilde{p} \mid V_1, \dots, V_n.$$

- **Remark**: many quantities of ecological interest are functionals of $\tilde{p} \implies$ this leads to natural Bayesian estimators for coverage, diversity, etc.

- Common nonparametric priors are the Dirichlet process (DP) and the Pitman–Yor (PY) process.

The discussion of Peter McCullagh (2016)



Two Early Contributions to the Ewens Saga

Peter McCullagh

Abstract. The mixture model devised by Fisher, Corbet and Williams [*Journal of Animal Ecology* **12** (1943) 42–58] for species sampling and the sequential prediction approach pioneered by Good [*Biometrika* **40** (1953) 237–264] and Good and Toulmin [*Biometrika* **43** (1956) 45–63] are both closely related to the Ewens sampling formula. Fisher's two-parameter joint distribution for the species counts includes the Ewens distribution as the conditional distribution given the sample size. The log-series model, as it is known in the ecological literature, is closely related to a Poisson process model devised by Arratia, Barbour and Tavaré [*Ann. Appl. Probab.* **2** (1992) 519–535]. Oddly, despite its advantages for statistical inference, Fisher does not mention the conditional distribution. Likewise, athough Good (1953) pioneered the sequential prediction approach, neither he nor Toulmin discovered the Ewens process in a form equivalent to the modern-day Chinese restaurant process.

Key words and phrases: Chinese restaurant process, Poisson process, species richness, species sampling.

The interplay between frequentist and Bayesian methods

- There is a clear and strong **interplay** between frequentist and Bayesian procedures. Using Peter McCullagh words

  > *"It is fair to say that Fisher almost discovered the Ewens sampling formula"*.

  The Ewens sampling formula is just another way of defining the Dirichlet process.

- **However**:

- The properties of frequentist estimators are often based on asymptotic considerations $\Longrightarrow$ Bayesian inference could be helpful.

- If **prior information** is available, there is not a simple way to incorporate it into the modeling.

- It is even more problematic to incorporate these estimators into **more complex models** e.g. accounting for covariates.

The Pitman–Yor model I

- Let us consider the **Pitman–Yor** process, written $\tilde{p} \sim \mathrm{PY}(\alpha, \sigma)$, with parameters $\sigma \in [0,1)$ and $\alpha > -\sigma$. When $\sigma = 0$ this reduces to the Dirichlet process.

- Marginalization over $\tilde{p}$ leads to specification for the joint distribution of $V_1, \dots, V_n$.

- This can be conveniently expressed through a **sequential mechanism**.

- In a PY model, the probability that the $(n+1)$st sequence belongs to the jth of the known taxa is

$$\mathrm{pr}(V_{n+1} = V_j^* \mid V_1, \dots, V_n) = \frac{n_j - \sigma}{\alpha + n}, \qquad j = 1, \dots, k,$$

while the probability of observing a new taxon is

$$\mathrm{pr}(V_{n+1} = \text{``new''} \mid V_1, \dots, V_n) = \frac{\alpha + \sigma k}{\alpha + n}.$$

- **Technical note**. The law of $\tilde{p}$ is fully characterized by the above sequential scheme, through **de Finetti representation** theorem.

The Pitman–Yor model II

Stick-breaking definition of the PY

$$\tilde{p} = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \qquad \pi_h = \nu_h \prod_{\ell=1}^{h} (1 - \nu_\ell), \quad \nu_h \overset{\text{ind}}{\sim} \text{BETA}(1 - \sigma, \alpha + \sigma h),$$

with $\sigma \in [0, 1)$ and $\alpha > -\sigma$.

Urn-scheme

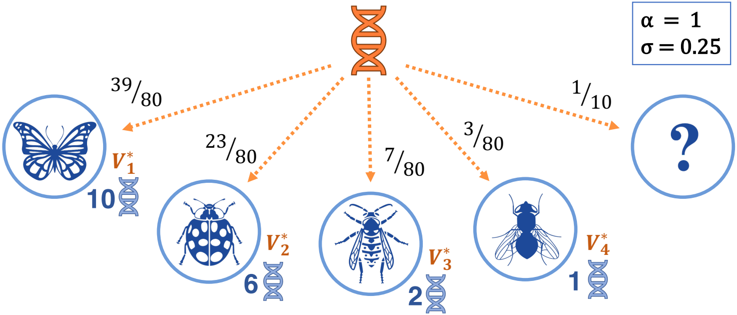
$$V_{n+1} \mid V_1, \dots, V_n \sim \frac{\alpha + \sigma k}{\alpha + n}(\text{"new species"}) + \frac{1}{\alpha + n} \sum_{j=1}^{k} (n_j - \sigma) \delta_{V_j^*}.$$

Posterior distribution

$$(\tilde{p} \mid V_1, \dots, V_n) = \sum_{j=1}^{k} W_j \delta_{V_j^*} + W_{k+1} \tilde{q},$$

with $(W_1, \dots, W_{k+1}) \sim \text{DIR}(n_1 - \sigma, \dots, n_k - \sigma, \alpha + \sigma K)$ and $\tilde{q}$ is a $\text{PY}(\alpha + \sigma k, \sigma)$.

Example: Pitman-Yor process



- Example of a Pitman–Yor process with $n = 19$, $\alpha = 1$, $\sigma = 0.25$ and $K_n = 4$.
- The probability of re-observing V_1^* is $(n_1 - \sigma)/(\alpha + n) = 39/80$.
- The probability for the new taxon is $(\alpha + \sigma k)/(\alpha + n) = 1/10$.

The discovery of new species

Proposition (Favaro et al., JRSS-B, 2009)

In a PY model a Bayesian estimate of the **rarefaction** curve is

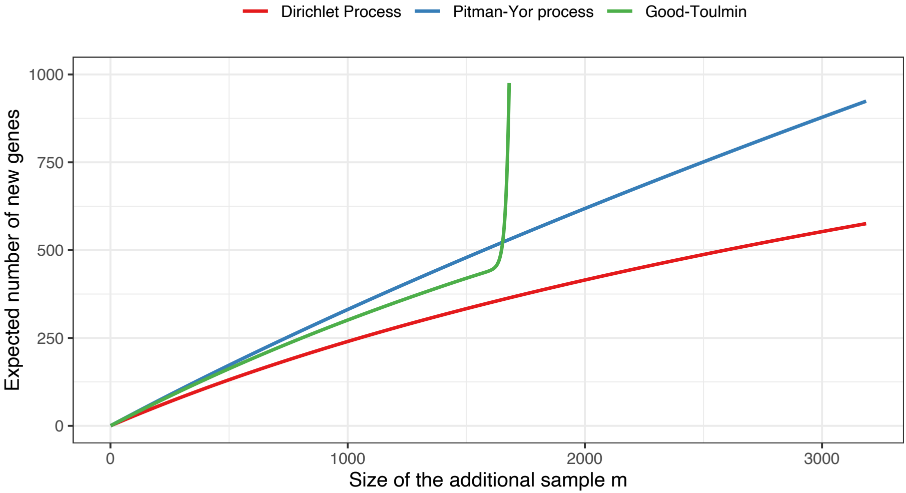
$$\mathbb{E}(K_n) = \frac{\alpha}{\sigma}\left\{\frac{(\alpha+\sigma)_n}{(\alpha)_n} - 1\right\}.$$

Moreover, Let $K_m^{(n)}$ be the number of new species we observe in an additional sample of size m. Then

$$\mathbb{E}(K_m^{(n)} \mid V_1, \dots, V_n) = \left(k + \frac{\alpha}{\sigma}\right)\left\{\frac{(\alpha+n+\sigma)_m}{(\alpha+n)_m} - 1\right\}.$$

- The parameters (α, σ) are often estimated using **maximum likelihood**.
- This Bayesian nonparametric setup has been developed by Lijoi et al. (2007), for the more general classes of Gibbs-type priors.

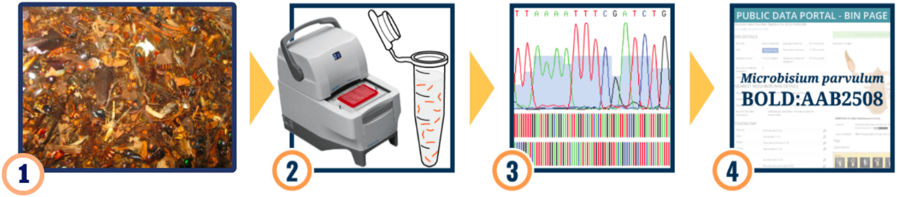
BNP prediction of the number of new species



Modern species discovery

- The models and techniques we just discussed are exciting from a theoretical perspective and are sometimes useful for getting preliminary estimates.
- Software: `https://alessandrozito.github.io/BNPvegan/vignette.html`
- The **exhangeability** assumption is unrealistic in many applied scenarios.
- Indeed, modern sampling strategies for species discoveries are much more sophisticated than those used in the '40...
- However, the **theoretical knowledge** of these processes in "controlled" settings enables the implementation of BNP ideas in much more complex scenarios.
- From here on, we make a big jump into a much more realistic application.

DNA barcoding


1
2
T T A A A A T T T C G A T C T G
3
PUBLIC DATA PORTAL - BIN PAGE
Microbisium parvulum
BOLD:AAB2508
4

- **DNA barcoding** is the practice of placing DNA sequences within a Linnean taxonomy.

- **Advantages**:
  - Processing and classification of a large number of query sequences in a reasonable time
  - No need for morphological identification, which is impossible with soups of insects

- **Challenges**:
  - Libraries of labeled DNA (reference libraries) are often incomplete
  - Many species are still unknown to science

Malaysian traps

A Malaysian trap

Statisticians "supervising" the data collection process

The data structure

- A **taxonomic library** is a collection of observations of the form $\mathcal{D}_n = (\mathbf{V}_i, \boldsymbol{X}_i)_{i=1}^n$.

- Each vector $\boldsymbol{V}_i = (V_{i,1}, \dots, V_{i,L})$ contains the **taxonomic annotations** for a DNA sequence. For example, we could have:

$$V_{i,1} = \text{“}\textit{Insecta}\text{”}, \qquad V_{i,2} = \text{“}\textit{Diptera}\text{”}, \qquad V_{i,3} = \text{“}\textit{Tephritidae}\text{”}, \qquad \text{etc.}$$

- Here L is the number of levels (layers) one wish to consider.

- Each $\boldsymbol{X}_i$ contains the **output** of the DNA **barcoding** procedure (e.g. the k-mer decomposition, or the aligned sequences).

- **Remark**: a relevant amount of pre-processing is needed to get $\boldsymbol{X}_i$.

- From a modelling perspective, the $\boldsymbol{V}_i$'s represent the **response variables** whereas $\boldsymbol{X}_i$'s correspond to the **covariates**.

A nested classification problem

- Given a new covariate $\boldsymbol{X}_{n+1}$ from the DNA barcoding procedure, we wish to **predict** the corresponding taxonomic labels $\boldsymbol{V}_{n+1}$.

- This is essentially a **classification problem**, albeit there are several statistical challenges $\implies$ off-the-shelves algorithms cannot be used without modifications.

- **Challenge 1**. The labels $\boldsymbol{V}_i$, by construction, have a **nested structure**. The method should be consistent with the taxonomy and possibly exploit it to improve accuracy.

- **Challenge 2**. We want a **probabilistic** and well-calibrated method to quantify the uncertainty associated with our predictions formally.

- **Challenge 3**. Some labels may be absent in the reference dataset or even be unknown to science. We need to account for the possibility that $V_{n+1,\ell} =$ "new".

- Moreover, both the training and the prediction step should be performed quickly, as the number of test covariates is huge.

Overview of the model

- Paralleling the construction, e.g. of **naive Bayes** classifiers and linear discriminant analysis, we will specify a joint distribution through the decomposition

$$p(\boldsymbol{V}^{(n+1)}, \boldsymbol{X}^{(n+1)}) = p(\boldsymbol{V}^{(n+1)})p(\boldsymbol{X}^{(n+1)} \mid \boldsymbol{V}^{(n+1)}),$$

where $\boldsymbol{V}^{(n+1)} = (\boldsymbol{V}_i)_{i=1}^{n+1}$ and $\boldsymbol{X}^{(n+1)} = (\boldsymbol{X}_i)_{i=1}^{n+1}$.

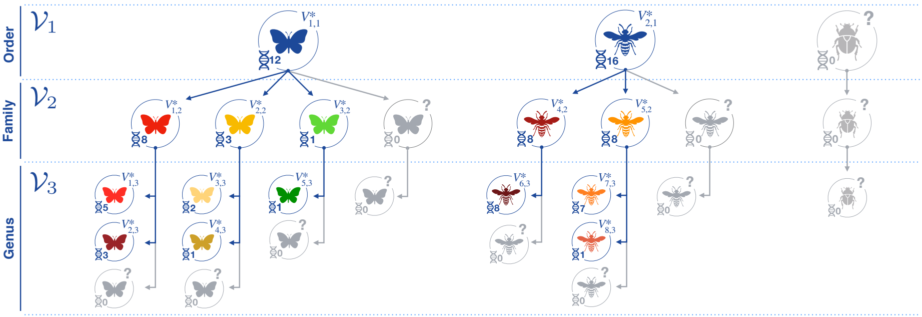
- Then, we will obtain a closed-form expression for the **predictive** distribution

$$p(\boldsymbol{V}_{n+1} \mid \boldsymbol{X}_{n+1}, \mathcal{D}_n) \propto p(\boldsymbol{V}_{n+1} \mid \boldsymbol{V}^{(n)})p(\boldsymbol{X}^{(n+1)} \mid \boldsymbol{V}^{(n+1)}),$$

recalling that $\mathcal{D}_n = (\mathbf{V}_i, \boldsymbol{X}_i)_{i=1}^{n}$, which can be used for taxonomic classification.

- Each $\boldsymbol{V}_i$ is discrete, therefore the normalizing constant is trivial to compute.

Prior specification for taxonomic labels



- In the first place, we need to specify a distribution for the taxonomic labels $p(\boldsymbol{V}^{(n+1)})$.

- The need for species discoveries (actually, taxon discoveries in this case) naturally calls for **Bayesian nonparametric** tools.

Enriched Pitman–Yor process ($L = 2$)

- The natural next step in the specification of $p(\boldsymbol{V}^{(n+1)})$ is the $L = 2$ case.

- A possibility is the usage of **enriched processes**, such as the enriched Dirichlet process of Wade et al. (2011) and subsequent generalizations to the Pitman–Yor case.

- When $\boldsymbol{V}_i = (V_{i,1}, V_{i,2})$, for the first taxon one could assume

$$(V_{i,1} \mid \tilde{p}_1) \overset{\text{iid}}{\sim} \tilde{p}_1, \quad \tilde{p}_1 \sim \mathrm{PY}(\alpha_1, \sigma_1).$$

- For the second taxon, instead, we let

$$(V_{i,2} \mid V_{i,1} = v, \tilde{p}_{2,v}) \overset{\text{iid}}{\sim} \tilde{p}_{2,v} \qquad \tilde{p}_{2,v} \overset{\text{ind}}{\sim} \mathrm{PY}\{\alpha_2(v), \sigma_2(v)\}$$

for $\alpha_2(v) > -\sigma_2(v)$, $\sigma_2(v) \in [0, 1)$.

- This induces the desired **nested** behaviour and induces a specification for $p(\boldsymbol{V}^{(n+1)})$ through the so-called **enriched urn-scheme**.

Taxonomic Pitman–Yor priors

- Enriched processes can be naturally extended to general **taxonomic priors**.
- At the first level, we let as before $V_{i,1} \mid \tilde{p}_1 \overset{iid}{\sim} \tilde{p}_1, \quad \tilde{p}_1 \sim \mathrm{PY}(\alpha_1, \sigma_1)$.
- For all the subsequent layers $\ell = 2, \dots, L$, we specify the following **nested** priors

$$(V_{i,\ell} \mid V_{i,\ell-1} = v, \tilde{p}_{\ell,v}) \overset{\text{iid}}{\sim} \tilde{p}_{\ell,v} \qquad \tilde{p}_{\ell,v} \overset{\text{iid}}{\sim} \mathrm{PY}(\alpha_\ell, \sigma_\ell).$$

Remark. To simplify our modeling, we let $\alpha_\ell(v) = \alpha_\ell$ and $\sigma_\ell(v) = \sigma_\ell$.

- This produces a **taxonomic urn scheme**, in which

$$(V_{\ell,n+1} \mid V_{\ell-1,n+1} = v, \mathbf{V}_{\ell,\cdot}^{(n)}) = \begin{cases} \text{``new''} & \{\alpha_\ell + \sigma_\ell K(v)\}/\{\alpha_\ell + n(v)\} \\ V_{\ell,j}^* & \{n(V_{\ell,j}^*) - \sigma_\ell\}/\{\alpha_\ell + n(v)\} \end{cases}$$

with $\mathbf{V}_{\ell,\cdot}^{(n)} = (V_{\ell,i})_{i=1}^n$, and where $n(v)$ and $K(v)$ are the number of sequences and distinct nodes linked to v.

Taxonomic Pitman–Yor priors

- Under a **taxonomic prior**, the probability of the future taxonomic label $\boldsymbol{V}_{n+1}$ can be obtained using **chain rule**.

- Indeed, the latter is a product of the Pitman–Yor probabilities associated to $\boldsymbol{v}$, so that

$$\begin{aligned}\mathrm{pr}(\boldsymbol{V}_{n+1} = \boldsymbol{v} \mid \mathbf{V}^{(n)}) &= \mathrm{pr}(\boldsymbol{V}_{n+1} = (v_1, \dots, v_L) \mid \mathbf{V}^{(n)}) \\ &= \mathrm{pr}(V_{n+1,1} = v_1 \mid \mathbf{V}^{(n)}) \prod_{\ell=2}^{L} \mathrm{pr}(V_{n+1,\ell} = v_\ell \mid V_{n+1,\ell-1} = v_{\ell-1}, \mathbf{V}^{(n)})\end{aligned}$$

where $\mathbf{V}^{(n)} = (\mathbf{V}_i)_{i=1}^{n}$ is the collection of all taxonomic labels up to n.

- This scheme describe a distribution for the taxonomic labels $p(\boldsymbol{V}^{(n+1)})$, as desired.

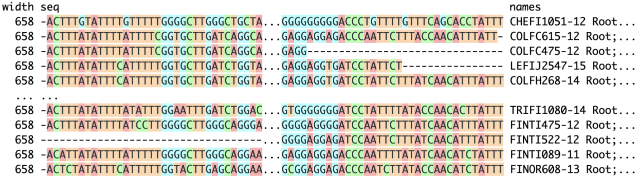
The likelihood component

- We let the output of the DNA sequence depend on the taxa as follows

$$(\boldsymbol{X}_i \mid \boldsymbol{V}_i = (v_1, \dots, v_L), \boldsymbol{\theta}_{v_L}) \overset{\text{ind}}{\sim} \mathcal{K}(x; \boldsymbol{\theta}_{v_L}),$$

  for some generic **kernel** $\mathcal{K}$, which must be selected according to the data structure.

- For example, let us assume the DNA sequences are **globally aligned**, so that:



```
    width seq                                                                             names
      658 -ACTTTGTATTTTGTTTTTGGGGCTTGGGCTGCTA...GGGGGGGGGACCCTGTTTTGTTTCAGCACCTATTT CHEFI1051-12 Root...
      658 -ACTTTATATTTTATTTTCGGTGCTTGATCAGGCA...GAGGAGGAGACCCAATTCTTTACCAACATTTATT- COLFC615-12 Root;...
      658 -ACTTTATATTTTATTTTCGGTGCTTGATCAGGCA...GAGG------------------------------- COLFC475-12 Root;...
      658 -ACTTTATATTTCATTTTTGGTGCTTGATCTGGTA...GAGGAGGTGATCCTATTCT---------------- LEFIJ2547-15 Root...
      658 -ACTTTATATTTCATTTTTGGTGCTTGATCTGGTA...GAGGAGGTGATCCTATTCTTTATCAACATTTATTT COLFH268-14 Root;...
      ... ...
      658 -ACTTTATATTTTATATTTGGAATTTGATCTGGAC...GTGGGGGGGATCCTATTTTATACCAACACTTATTT TRIFI1080-14 Root...
      658 -ACTTTATATTTTATCCTTGGGGCTTGGGCAGGGA...GGGGAGGGGATCCAATTCTTTATCAACATTTATTT FINTI475-12 Root;...
      658 -----------------------------------...GGGGAGGAGATCCAATTCTTTATCAACATTTATTT FINTI522-12 Root;...
      658 -ACATTATATTTTATTTTTGGGGCTTGGGCAGGAA...GAGGAGGAGACCCAATTTTATATCAACATCTATTT FINTI089-11 Root;...
      658 -ACTCTATATTTCATTTTTGGTACTTGAGCAGGAA...GCGGAGGAGACCCAATCTTATACCAACATCTATTT FINOR608-13 Root;...
```

- Then, the nucleotides $\boldsymbol{X}_i = (X_{ij})_{j=1}^{p}$, $X_{ij} \in \{A, C, G, T, -\}$ can be modelled as a **product multinomial kernel**, namely

$$\mathcal{K}(x; \boldsymbol{\theta}_{v_L}) = \prod_{j=1}^{p} \prod_{g \in \{\text{A,C,G,T,-}\}} \theta_{v_L,j,g}^{\mathbb{1}\{x=g\}}, \quad \theta_{v_L,j} \sim \text{Dir}(\xi_{v_L,j,A}, \dots, \xi_{v_L,j,T}).$$

One step-ahead predictions

- The **predictive probability** of a new taxonomic label $\boldsymbol{V}^{(n+1)}$ is

$$\begin{aligned}\mathrm{pr}(\boldsymbol{V}_{n+1} = \boldsymbol{v} \mid \boldsymbol{X}_{n+1}, \mathcal{D}_n) &\propto \mathrm{pr}(\boldsymbol{V}_{n+1} = \boldsymbol{v} \mid \mathbf{V}^{(n)}) p(\boldsymbol{X}^{(n+1)} \mid \boldsymbol{V}^{(n+1)}) \\ &\propto \mathrm{pr}(\boldsymbol{V}_{n+1} = \boldsymbol{v} \mid \mathbf{V}^{(n)}) \int \mathcal{K}(\boldsymbol{X}_{n+1}; \boldsymbol{\theta}_{v_L}) p(\boldsymbol{\theta}_{v_L} \mid \mathcal{D}_n) \mathrm{d}\boldsymbol{\theta}_{v_L}.\end{aligned}$$

- The above integral is available in closed form under our specifications. Once the **hyperparameters** have been selected, this leads to a remarkably fast procedure.

- Note that $p(\boldsymbol{\theta}_{v_L} \mid \mathcal{D}_n)$ corresponds to the prior $p(\boldsymbol{\theta}_{v_L})$ if v_L is “new”.

- The probabilities of higher-level taxa can be obtained through marginalization in the elements of $\boldsymbol{V}_{n+1}$, which corresponds to summation here.

- **Classification rule**: iteratively select the taxon with the highest probability given the previously selected branch, preserving a meaningful taxonomic structure.

Further practical considerations

- The parameters $\boldsymbol{\theta}_{v_L}$ are specific to the last layer of the taxonomy and play an important role in novel species recognition.

- Hence, in order to **borrow strenght** across branches, the hyperparameters $\boldsymbol{\xi}_v$ have been estimated in a careful manner using a **method of moments** algorithm.

- As for the set of hyperparameters α_ℓ, σ_ℓ, they have been estimated via **empirical Bayes**, following standard practice.

- Finally, to ensure proper calibration, the predictive probabilities have been post-processed to account for model **misspecification**.

- Specifically, the predictive probabilities have been raised to a power $\rho \in (0, 1)$ and then renormalized. Optimal ρ is selected on a hold-out dataset.

Application: FinBOL insects dataset

- The Finland Barcode of Life initiative FinBOL data contains **34624 sequences** of **insect species** recorded in Finland, with a taxonomy of **seven levels**.

- Seven levels: *Class, Order, Family, Subfamily, Tribe, Genus, Species* - $10,985$ distinct *Species*.

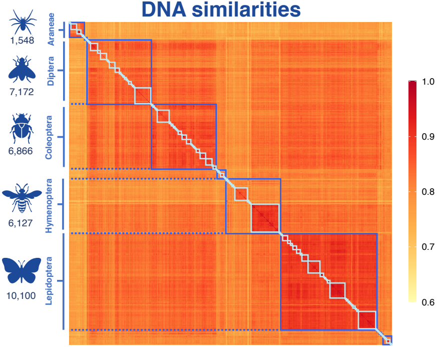
- How does our model work under **incomplete libraries**?

- **Our strategy**: employ different train-test splitting strategies
  - **Split on Sequences**: purely random subset. Each query has an equal probability of being on the test.
  - **Split on Taxa**: stratified random subset. First select a taxon, and then a sequence within that taxon.

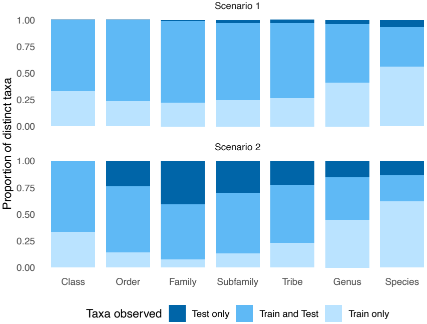
- In the first split train and test are "similar". In the second they have different compositions.

Application: FinBOL insects dataset

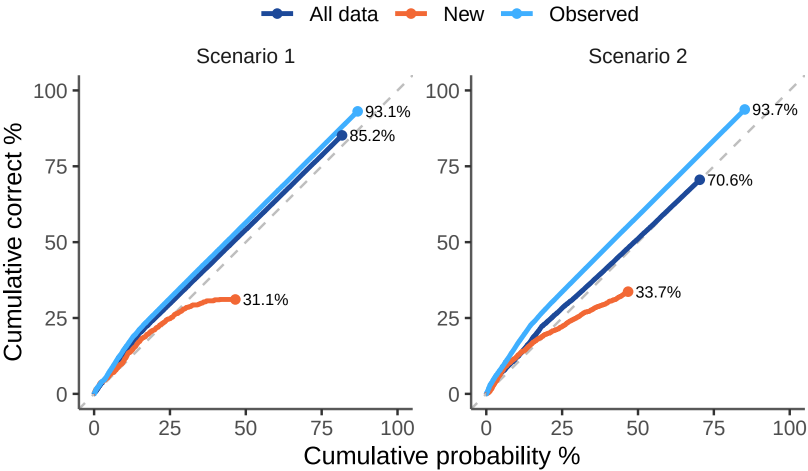
DNA similarities



Train-test composition



Calibration plot



Test set accuracy

- Results on the FinBOL dataset, on a hold-out dataset (split on sequences).

MODEL	CLASS	ORDER	FAMILY	SUBFAMILY	TRIBE	GENUS	SPECIES
M-1	100.0	99.9	98.6	97.5	96.0	92.1	85.2
	(1)	(1)	(.98)	(.96)	(.94)	(.91)	(.82)
M-2	100.0	99.9	98.4	97.2	95.8	92.4	85.4
	(1)	(1)	(.98)	(.97)	(.95)	(.93)	(.86)

- Models M-1 and M-2 refer to two distinct specifications for the kernel.

- Values report the **percentage** of DNA sequences **correctly labelled**.

- Values in parenthesis denote the **average prediction probabilities** in the test set.

Summary

- The proposed model **BayesANT** is a **probabilistic** taxonomic classifier that accounts for the possibility of observing new species.
- It can be regarded as a **covariate-dependent** species sampling model.
- The modeling choices reflect a balance between **flexibility** and **pragmatism** in developing an efficient algorithm that can easily handle millions of sequences.
- Crucially, this is because closed-form expressions for the predictive structure are available.
- Software: `https://alessandrozito.github.io/BayesANT/vignette.html`

Main reference

Zito, A., Rigon, T. and D. B. Dunson (2023). Inferring Taxonomic Placement from DNA Barcoding Allowing Discovery of New Taxa. *Methods in Ecology and Evolution* **14**: 529–42.