MCEB 2024

(https://mceb2024.sciencesconf.org/?lang=en)

Le Hameau de l'Etoile[1]
17-21 June

Monday, June 17

18:00 Welcoming of participants. Apéritif

20:00 Dinner

Tuesday, June 18

09:00-10:00 (Plenary address) **Tal Pupko**

10:00-10:30 Break

10:30-10:50
Kate Truman "*Identifiability of the Fossilised Birth-Death model*"

10:50-11:10
François Rousset. "*Better confidence intervals in simulation-based inference*"

[1] https://maps.app.goo.gl/gMAnSEfvF2boZt1Q6

11:10-11:30
Miguel De Navascués. "*Demographic inference from radiocarbon data*"

11:40-12:00
Andreas Futschik. "Detecting selection using signature kernels: A novel method for multi-locus Wright-Fisher models with recombination"

12:00-12:20
Anna Zhukova "*Epidemiological birth-death models with partner notification*"

12:30-14:00 Lunch break

14:00-17:30 Discussions

17:30-18:30 (Plenary address) **Claire Guinat**

18:30-20:00 Apéritif and poster session 1

20:00 Dinner

Wednesday, June 19

09:00-10:00 (Plenary address) **Claudia Solis-Lemus**

10:00-10:30 Break

10:30-10:50
Scott Edwards "*New Bayesian methods for linking genomic and phenotypic variation*"

10:50-11:10
Isabel San Martín. "*Assessing the power of artificial intelligence approaches for birth-death models*"

11:10-11:30

Antoine Aragon "*Learning evolutionary parameters from allelic trees*"

11:40-12:00
Simon Boitard "*SelNeTime : a new method inferring demography and selection from genomic time series data*"

11:00-12:20
Cécile Ané "*Identifying circular orders for blobs in phylogenetic networks*"

12:30-14:00 Lunch break

Free afternoon

20:00 Dinner

Thursday, June 20

09:00-10:00 (Plenary address) **Jean-Michel Marin**

10:00-10:30 Break

10:30-10:50
Marius Brusselmans "*On the importance of assessing topological convergence in Bayesian phylogenetic inference*"

10:50-11:10
Luc Blassel "*Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks*"

11:10-11:30
Katharina Huber "*Ploidy profiles and phylogenetic networks with horizontal arcs*"

11:40-12:00
Carlos Albors "*Substitution rate prediction as a pretraining task for models of human genomes*"

12:00-12:20
Marta Pelizzola "*Robust learning of mutational signatures using non-negative matrix factorization*"

12:30-14:00 Lunch break

14:00-17:30 Discussions

17:30-18:30 (Plenary talk) **Yun S. Song**

18:30-20:00 Apéritif and poster session 2

20:00 Dinner

Friday, June 21

09:00-10:00 (Plenary address) **Caroline Colijn**

10:00-10:30 Break

10:30-10:50
Wakinyan Benhamou "*Inferring the life-history traits of new viral variants from epidemiological and evolutionary dynamics*"

10:50-11:10
Lars Berling "*Statistics in the space of ranked time trees*"

11:10-11:30
Frédéric Lemoine "*The Bayesian phylogenetic bootstrap and its application to short branches and trees*"

11:30-11:50
Hector Banos "*Topics in profile mixture models: Overparameterization and a Linked general-time reversible model*"

12:50-12:10
Carolin Kosiol "*Polymorphism-aware models in RevBayes: Species trees, disentangling Balancing Selection and CG-biased gene conversion*"

12:30-13:30 Lunch break

13:30 Departure to Montpellier (arrival scheduled at 14:30 at Gare Routière)

Poster session 1

- Jordan Moutet "*Algorithms to reconstruct past indels: a parsimony approach*"
- Amélie Ngo "*Applying mathematical models to unravel pathogen evolution: identifying severe phenotype association through computational phylogenetics*"
- Valentino Giulio Dalla Riva "*Are evolutionary unique species ecologically unique?*"
- Yuki Takasawa "*Beyond the majority rule consensus tree*"
- Mael Guivarch "*Comparing clustering methods to highlight fine-scale genetic structure using simulations and POPGEN data*"
- Bastien Boussau "*Detection of site-specific selection regimes in protein alignments using transformers*"

Poster session 2

- Anastasis Togkousidis "*Early stopping in Maximum Likelihood Phylogenetic Inference*"
- Askar Gafurov "*Efficient analysis of annotation colocalization accounting for genomic contexts*"
- Nikita Kulikov "*Machine learning phylogenetics empowered by dimensionality reduction techniques*"
- Harry Gordon "*Simulation based inference of the evolutionary history of wildcats (Felis silvestris): Machine learning for population genomics*"
- Vincent Moulton "The space of equidistant cactuses"
- Daniel Hudson "Trees, networks, QR codes and the SplitsTree App"

Identifiability of the Fossilised Birth-Death Model

Kate Truman **1** , Timothy Vaughan **2** , Alex Gavryushkin **1** , Alexandra Gavryushkina **1**
1 : University of Canterbury
Christchurch - New Zealand
2 : Department of Biosystems Science and Engineering [ETH Zürich]

Time-dependent birth-death sampling models have been used in numerous studies for inferring past evolutionary dynamics in different areas, e.g. speciation and extinction rates in macroevolutionary studies, or effective reproductive number in epidemiological studies. These models are branching processes where lineages can bifurcate, die, or be sampled with time-dependent birth, death, and sampling rates and generate phylogenetic trees. It has been recently shown that in some subclasses of such models, different sets of rates can result in the same distributions of reconstructed phylogenetic trees, and therefore the rates become unidentifiable from the trees regardless of their size. Here I present the finding that widely used time-dependent fossilised birth-death (FBD) models are identifiable. This subclass of models makes more realistic assumptions about the fossilisation process and certain infectious disease transmission processes than the unidentifiable birth-death sampling models. Namely, FBD models assume that sampled lineages stay in the process rather than being immediately removed upon sampling. Identifiability of the time-dependent FBD model ensures that statistical methods that implement this model infer the true underlying temporal diversification or epidemiological dynamics from phylogenetic trees or directly from molecular or other comparative data. I will further cover that the time-dependent birth-death model with an extra parameter, the removal after sampling probability, is unidentifiable. This implies that in scenarios where we do not know how sampling affects lineages we are unable to infer this extra parameter together with birth, death, and sampling rates solely from trees.

Better confidence intervals in simulation-based inference

François Rousset **1,@** , Raphael Leblois, Arnaud Estoup, @
1 : Institut des Sciences de l'Evolution de Montpellier
Centre National de la Recherche Scientifique

In this work we describe and evaluate a method of statistical inference revisiting the idea of inferring a likelihood surface by simulation when the likelihood function cannot be evaluated. The method aims in particular to provide confidence intervals with controlled

coverage and its performance is assessed accordingly. It is compared un particular to that of approximate Bayesian computation with random forests (ABC-RF), with which it shares some technical features, for equivalent simulation effort of the studied biological processon scenarios of historical population divergence and admixture. The comparison highlights the good performance of the new method, and the importance of an iterative workflow for exploring the parameter space efficiently.

Demographic inference from radiocarbon data

Miguel De Navascués **1, 2, @**

1 : CBGP, INRAE, CIRAD, IRD, Institute Agro, University of Montpellier, Montpellier, France

CBGP, INRAE, CIRAD, IRD, Institute Agro, University of Montpellier, Montpellier, France

2 : Human Evolution, Uppsala Universitet, Sweden

The development of radiocarbon dating revolutionized the study of the past, with its applications in archaeology and paleobiology. As the technique became a standard in research, the accumulation of dated samples leaded to the study of the abundance of these samples through time to study changes in population size in humans and other species. However, the classical analysis of the abundance of radiocarbon data relies in a visual evaluation the estimated abundance of samples. Here I propose an approximate Bayesian computation approach that allows to perform model choice and parameter estimates. The proposed model takes into account the uncertainty of radiocarbon age. I will discuss the potential to combine this data with other type of data, such as population genetic data, to do demographic inference.

Detecting Selection using Signature Kernels: A Novel Method for Multi-Locus Wright-Fisher Models with Recombination

Yuehao Xu **1** , Sherman Khoo **2** , Andreas Futschik **1, @** , Ritabrata Dutta **2**

1 : Johannes Kepler Universität Linz

2 : Department of Statistics [Warwick]

We propose an innovative Bayesian framework tailored for the inference of the selection coefficients in multi-locus Wright-Fisher models. Utilizing a signature kernel score, our approach offers an innovative solution for approximating likelihoods by extracting informative signatures from the trajectories of haplotype frequencies. Moreover, within a generalized Bayesian posterior framework, we derive the scoring rule posterior, which we then pair with a Population Monte Carlo (PMC) algorithm to obtain posterior samples for selection coefficients. This powerful combination enables us to infer selection dynamics efficiently even in complex high-dimensional and temporal data settings. Our method works well through extensive tests on both simulated and real-world data. Notably, our approach effectively detects selection both in univariate, and multivariate Wright-Fisher models, including 2-locus and 3-locus models with recombination. Our proposed novel technique contributes to a better understanding of complex evolutionary dynamics.

Epidemiological birth-death models with partner notification

Anna Zhukova **1, 2,** @ , Olivier Gascuel, @

1 : Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris
Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris
2 : G5 Evolutionary Dynamics of Infectious Diseases
Institut Pasteur de Paris, Université de Paris

Multi-type birth-death (MTBD) models have been useful for epidemiological parameter estimation from pathogen sequence data. However, some of the assumptions behind these models are unrealistic. For instance, these models assume that infected individuals stop being infectious at a constant rate (e.g. due to diagnostic and treatment), upon which their pathogen might get sampled (based on a probability given by the model). In practice, sampling is often non-random. In particular, upon diagnosis some of the patient's contacts (e.g. sexual partners in case of HIV) might be identified and offered testing (notified).
We introduce an extension of MTBD models that accounts for potential sampling due to partner notification, MTBD-PN. We propose a test for detecting partner notification in pathogen phylogenetic trees. For the simplest representative of this model family, the BD-PN model, we also present its parameter (e.g. R0) estimator. We test our estimator on simulated data, and apply it to the study of HIV-1 B epidemic in the UK.

Assessing the power of artificial intelligence approaches for birth-death models

Pablo Gutiérrez De La Peña **1,@** , Guillermo Iglesias Hernández **2** , Andrea Sánchez Meseguer **1** , Isabel Sanmartín **1**

1 : Real Jardín Botánico - CSIC

2 : Universidad Politécnica de Madrid

Birth-death (BD) models applied to dated phylogenies are a useful tool to study past diversification dynamics in the absence of complete fossil record. Parameters in these stochastic models are typically inferred using likelihood-based methods such as Maximum Likelihood or Bayesian Inference. However, these methods require the formulation of a new likelihood algorithm each time a new model is proposed, and some of the most complex models are also computationally intractable, or mathematically non-identifiable. The last years have witnessed a revolution, with artificial intelligence (AI) methods applied to phylogenetic inference, species delimitation, or epidemiology. However, the power of these approaches in birth-death modeling remains virtually unexplored. Here, we address a classification and regression problem, the power of AI algorithms to discriminate among six different diversification scenarios: constant birth–death (BD), high extinction (HE), mass extinction (ME), stasis and radiate (SR), diversity dependent (DD), and waxing and waning (WW). We simulated 100,000 trees under each diversification scenario and encoded the phylogenies using two different representation techniques: a set of summary statistics and a vector encoding. These simulations were used to train and validate two different AI methods: convolutional neural networks (CNN) and random forests (RF), and the trained AI algorithm was used to predict the most probable model of diversification for selected empirical phylogenies. Finally, we compared the performance of AI methods with previous likelihood-based approaches. The cross-validation approach showed that the percentage of correct assignments for the simulated scenarios was higher for the two DL algorithms than for likelihood-based approaches with similar size phylogenies. The most accurate strategy was the combination of CNN with the tree vectorial representation. We obtained similar levels of accuracy in the estimation of model parameters among the different strategies.

`

Learning evolutionary parameters from allelic trees

Antoine Aragon **1,@** , Thierry Mora **2** , Aleksandra Walczak **1** , Amaury Lambert **3**

1 : Laboratoire de physique de l'ENS - ENS Paris
Centre National de la Recherche Scientifique, Département de Physique de l'ENS-PSL
2 : Laboratoire de physique de l'ENS - ENS Paris
Département de Physique de l'ENS-PSL, Centre National de la Recherche Scientifique - CNRS
3 : Institut de biologie de l'ENS Paris
Département de Biologie - ENS Paris, Centre National de la Recherche Scientifique

Characterising selection and its impact on genetic diversity within evolving populations is a fundamental challenge in evolutionary biology. B-cell affinity maturation, a process exemplifying accelerated evolution under selection, offers a unique opportunity to investigate these dynamics within shorter timescales. Despite its significance, current methods face limitations: they often do not extend beyond neutrality, and there is a lack of reliable summary statistics for discerning non-neutral evolutionary processes. To address this issue, we construct mathematical models that capture the dynamics of proliferating cell populations undergoing non-neutral mutations, and, informed by these models we develop an expectation maximisation inference scheme to learn the parameters of evolution from data. We will show progress towards applying the algorithm to B-cell repertoire data.

SelNeTime : a new method inferring demography and selection from genomic time series data

Simon Boitard **1,@** , Mathieu Uhl **1, 2** , Miguel De Navascués **1** , Bertrand Servin **3**

1 : Centre de Biologie pour la Gestion des Populations
Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut de Recherche pour le Développement, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Institut Agro Montpellier, Université de Montpellier
2 : Institut des Sciences de l'Evolution de Montpellier
Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Ecole Pratique des Hautes Etudes, Institut de recherche pour le développement [IRD] : UR226, Centre National de la Recherche Scientifique, Université

de Montpellier, Centre de Coopération Internationale en Recherche Agronomique pour le Développement : UMR116, Centre National de la Recherche Scientifique : UMR5554
3 : Génétique Physiologie et Systèmes d'Elevage
Ecole Nationale Vétérinaire de Toulouse, École nationale supérieure agronomique de Toulouse, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Ecole Nationale Supérieure Agronomique de Toulouse

Genomic samples collected for a same population at several generations provide direct access to the genetic diversity changes occurring within a specific time period, informing us about both the demographic and adaptive processes acting on the population during that period. A common approach to analyze such data is to model observed allele frequencies using a Hidden Markov model (HMM) ; this approach allows computing the full likelihood of the data, while accounting both for the stochastic evolution of population allele frequencies along time and for the noise arising from sampling a limited number of individuals at each observed generation. Several such HMM methods have been proposed so far, differing mainly in the way they model the transition probabilities of the Markov chain. Following Paris et al (2019), we consider here the Beta with Spikes approximation, which avoids the computational issues associated to the Wright-Fisher model while still including fixation probabilities, in contrast to other standard approximations of this model like the Gaussian or Beta distributions. To enhance the potential of genomic time series data, we present an improved version of Paris et al (2019)'s approach, denoted SelNeTime, whose computation time is drastically reduced and which accurately estimates effective population size in addition to the selection intensity at each locus. As an illustration, we apply this approach to genome-wide data produced by an Evolve & Resequence experiment in the pest insect D. suzukii.

Identifying circular orders for blobs in phylogenetic networks

Cécile Ané **1,@** , John Rhodes **2** , Hector Baños **3** , Jingcheng Xu **1**
1 : University of Wisconsin-Madison
2 : University of Alaska Fairbanks
3 : California State University [San Bernardino]

Phylogenetic networks and admixture graphs allow for the representation and modelling of gene flow, introgression and hybridization. Under the assumption that the network is of level-1, most of its topology and edge parameters are known to be

identifiable, from various data types and under various models. But little is known about what network features are identifiable if the level-1 constraint is removed.
I will focus here on outer-labeled planar networks, and show that (1) the circular order of taxon blocks is unique for such networks around each blob, (2) the circular order information from 4-taxon subnetworks identifies the full circular order of each blob, and (3) this circular order information is identifiable from many data types and under various models, including models with incomplete lineage sorting. This is the first detailed general result on identifiability of network topology beyond the level-1 case. I will also give examples of different blob topologies that cannot be distinguished under many models and data types.

On the importance of assessing topological convergence in Bayesian phylogenetic inference

Marius Brusselmans **1,@** , Luiz Max Carvalho **2** , Samuel Hong **1** , Jiansi Gao **3** , Frederick Masten Iv **3** , Andrew Rambaut **4** , Philippe Lemey **1** , Marc Suchard **5** , Gytis Dudas **6** , Guy Baele **1**
1 : Catholic University of Leuven - Katholieke Universiteit Leuven
2 : Fundação Getulio Vargas - Escola de Matemática Aplicada [Rio de Janeiro]
3 : Fred Hutchinson Cancer Research Centre
4 : University of Edinburgh
5 : University of California
6 : Vilnius University [Vilnius]

Modern phylodynamics is often performed within a Bayesian framework, using sampling algorithms such as Markov chain Monte Carlo (MCMC) to approximate the posterior distribution. These algorithms require careful evaluation of the quality of the generated samples before inferences can be made. Within the field of phylogenetics, one frequently adopted diagnostic approach is to evaluate the effective sample size (ESS) and to investigate trace graphs of the sampled parameters. A major limitation of these approaches is that they are developed for continuous parameters and therefore incompatible with a crucial parameter in these inferences: the topology of the phylogenetic tree. Several recent advancements have aimed at extending these diagnostics to topological space. We present a case study illustrating how these topological diagnostics can contain information not necessarily found in standard diagnostics, and how decisions regarding which of these diagnostics to compute (and how) can impact inferences regarding MCMC convergence and mixing. Given the major importance of detecting convergence and mixing issues in Bayesian phylogenetic analyses, the lack of a unified approach to this problem warrants further action, especially now that additional tools are becoming available to researchers.

Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks

Luca Nesterenko [1] , Luc Blassel [1,@] , Laurent Jacob [2] , Bastien Boussau [1]
1 : Département écologie évolutive [LBBE]
Laboratoire de Biométrie et Biologie Evolutive - UMR 5558
2 : Biologie Computationnelle et Quantitative = Laboratory of Computational and Quantitative Biology
Sorbonne Université, Centre National de la Recherche Scientifique, Institut de Biologie Paris Seine

The reconstruction of a phylogenetic tree from a multiple sequence alignment is a crucial step in numerous bioinformatic pipelines. Unfortunately, state-of-the-art methods for phylogenetic reconstruction, namely Maximum likelihood and Bayesian inference, have a high computational cost, which limits their usability on large datasets and their applicability when complex models of evolution are used. The application of deep learning methods still remains underdeveloped in the field, with attempts that mostly have been limited to the reconstruction of quartet trees, addressing phylogenetic reconstruction as a classification problem and thus facing the issue of the super-exponential growth of the number of possible tree topologies with the number of sequences. We present a different supervised learning approach with a transformer-based network architecture that, given a multiple sequence alignment, predicts all the pairwise evolutionary distances between the sequences, which in turn allow us to accurately reconstruct the tree topology with existing distance-based algorithms. The neural network architecture and its high degree of parameter sharing allow us to apply the same network to alignments of arbitrary size, both in the number of sequences and in their length. Through a variety of simulation settings we show that the method reaches state-of-the-art accuracy in terms of distance and branch length prediction. In terms of topological accuracy, we observe that it consistently outperforms existing distance-based approaches. In some conditions, it even approaches Maximum Likelihood performance, while being significantly faster and not limited by the complexity of the underlying model of evolution. On data simulated with more realistic models of evolution that take into account coevolution between positions, or on data simulated with different types of selective pressure, the approach outperforms all existing methods. Overall, our results showcase the potential of our method to make phylogenetic inference tractable under models of sequence evolution that are too costly to run for state-of-the-art inference methods.

Ploidy profiles and phylogenetic networks with horizontal arcs

Katharina Huber **1, @**
1 : University of East Anglia [Norwich]

Polyploidisation is an evolutionary phenomenon by which an organism acquires multiple copies of its complete set of chromosomes. This number is sometimes called the *ploidy level* of a species. The main strategy for reconstructing the evolutionary past of a dataset that has undergone polyploidization is to first construct a so called multiple-labelled tree from the dataset and to then somehow derive a phylogenetic network from it that explains the dataset's *ploidy profil*e (i.e. the ploidy levels of the species that make up the dataset). The following question therefore arises: How much can be said about that past if such a tree is not readily available? In this talk, we first formalize this question and then present a novel link between ploidy profiles and phylogenetic networks with horizontal arcs. We conclude with a brief discussion of some open problems. This also includes problems that might be addressable using Machine Learning approaches.

Substitution Rate Prediction as a Pretraining Task for Models of Human Genomes.

Carlos Albors, Yun S. Song

UC Berkeley

Neural networks trained to predict the effects of genetic variants have enabled the interpretation of genomic data in silico with state-of-the-art accuracy. However, the intractability of acquiring more training data makes it challenging to improve their performance. A solution is to use transfer learning methods to leverage data that are indirectly related to tasks of interest. To this end, we introduce a task that utilizes data on base pair evolution to infer genomic constraints imposed by mutation and selection. We show that pretraining on this task improves the performance of models trained to predict functional annotations from sequence context.

Robust learning of mutational signatures using non-negative matrix factorization

Marta Pelizzola **1,@** , Ragnhild Laursen **1** , Asger Hobolth **1**
1 : Department of Mathematics (Aarhus University)

Mutational signatures are probability vectors over the different mutation types that represent specific mutational processes. In cancer genomics, the mutational profile of a patient is a mixture of such mutational processes. Mutational signatures are usually derived using non-negative matrix factorization. To extract the mutational signatures we have to choose an error model for the observed mutational counts, which determines the underlying distributional assumption of the data. In most applications, the mutational counts are assumed to be Poisson distributed, but this is often overdispersed and leads to an overestimation of the number of signatures. We introduce a different error model where the mutational counts follow a Negative Binomial distribution and an approach to accurately estimate the number of signatures. Simulations are used to show that correctly specifying the error model and the number of signatures are crucial steps to identify robust signatures. Furthermore, we extend the definition of a mutation type and include flanking nucleotides further away from the base substitution and we show that this provides more robust and interpretable signatures that also increase the predictive power for unknown data.

`

Inferring the life-history traits of new viral variants from epidemiological and evolutionary dynamics

Wakinyan Benhamou **1,@** , François Blanquart, Marc Choisy, Thomas W. Berngruber, Sébastien Lion, Rémi Choquet, Sylvain Gandon, @
1 : Centre d'Ecologie Fonctionnelle et Evolutive
Ecole Pratique des Hautes Etudes, Centre National de la Recherche Scientifique, Université de Montpellier

Evolutionary epidemiology theory can help understand the time-varying selection acting on new pathogen variants. In particular, the strength of selection acting on more transmissible and/or more virulent variants is expected to change with the availability of

susceptible hosts in the population. This theoretical framework is adequate to infer the life-history traits of new variants during epidemics in various situations. Here, we show how to combine the analysis of the evolutionary dynamics with the epidemiological dynamics to infer key epidemiological quantities of new variants. First, we use data from an evolution experiment carried out with bacteriophage lambda in a chemostat to estimate the life-history traits of a specific mutation in the viral gene cI that controls the switch between lysis and lysogeny. We discuss how the experimental protocol could be optimized to improve our ability to estimate some life-history traits. Second, we show that we can use similar approaches at a very different scale to infer the lifehistory traits of human pathogens. More specifically, we show how to estimate the increase in the transmission rate or in the duration of infectiousness from an analysis of the spread of the Alpha variant in the early stage of the SARS-CoV-2 pandemic.

Statistics in the space of ranked time trees

Lars Berling **1,@** , Alex Gavryushkin **1**
1 : University of Canterbury
Christchurch - New Zealand

Phylogenetic trees form high dimensional non-Eulidean spaces, rendering classical statistical approaches ineffective and adding complexity to the problem. Estimating basic statistics, such as mean and variance for samples of trees is already challenging and using tree based statistics is mostly overlooked in practice. When evaluating sets of trees, it is common to use heuristics to compute a summary tree or to project the tree to a lower dimensional Euclidean vector space to apply standard statistics. However, few tools exist to analyse trees within their respective treespace and the most prominent candidate treespaces have been shown to exhibit unwanted properties. For example the 'stickiness' property of stratified spaces such as BHV or the computational complexity of tree-rearrangement based spaces such as NNI. Moreover, it has been shown that if one is interested in rooted time-trees the underlying geometry of these treespaces is fundamentally different. Here we introduce a recently developed treespace based on local tree rearrangements that allows for computational efficient distance computation, the ranked NNI space. We highlight desirable properties of this treespace that we believe are important for developing statistics within it. Additionally, we showcase applications such as estimating means for summarizing sets of trees and using tree variances in the context of MCMC convergence assessment.

The Bayesian Phylogenetic Bootstrap and its Application to Short Branches and Trees

Lemoine Frédéric **1,@** , Olivier Gascuel **2**

1 : Institut Pasteur [Paris]

Institut Pasteur de Paris

2 : ISYEB

Museum National d'Histoire Naturelle - MNHN (FRANCE)

Felsenstein's bootstrap is the most commonly used method to measure branch support in phylogenetics. Current sequencing technologies can lead to massive sampling of taxa or strains (as with SARS-CoV-2 responsible for the COVID-19 epidemic). In this case, many sequences are identical or nearly identical, and tree branches are often of length 0 or nearly 0 (corresponding to very few or no mutations). Nevertheless, these trees contain a strong signal, with unresolved portions but a low rate of incorrect branches. We demonstrate that with such data, Felsenstein's bootstrap requires adaptations. In particular, a branch of length zero (thus not supported by any mutation) which should have a non-significant bootstrap support, is sometimes assigned a high support. We propose simple adaptations to correct this effect. In addition, due to the frequentist nature of bootstrap sampling, the expected support of a branch corresponding to a single mutation is much less than 1.0, even if it is highly likely to be correct. To correct for this effect, we propose a Bayesian version of the phylogenetic bootstrap, in which sites, instead of being sampled with replacement as in the classical approach, are assigned uninformative prior probabilities. The branch support can then be interpreted as a posterior probability. This paradigm shift is consistent with the difference between frequentist and Bayesian approaches. We do not view the alignment as a small subsample (e.g., one gene) of a genome-sized sample of sites, but rather as containing all available information, which is biologically consistent if the data actually correspond to complete (e.g., virus) genomes or a large fraction of them. Unlike the frequentist approach, where bootstrap samples contain less information than the original alignment, a Bayesian sample still contains all the information without leaving some of the informative sites on the side. We give formulas for the expected support under the assumption of perfect phylogeny, in both the frequentist and Bayesian frameworks. Simulation results show that these theoretical results are robust against realistic data. Our approach is tested on simulated data and applied to SARS-CoV-2 and Ebola datasets. We show that the combination of considering null branches and using the Bayesian bootstrap generally removes falsely supported very short branches and makes the support of other branches more interpretable, with high support for correct branches.

Topics in profile mixture models: Overparameterization and a Linked general-time reversible model

Hector Banos **1,@** , Edward Susko **2** , Andrew J. Roger **2**
1 : California State University [San Bernardino]
2 : Dalhousie University [Halifax]

Phylogenetic models of protein sequence evolution not accounting for site heterogeneity are prone to long-branch attraction (LBA) artifacts, especially when reconstructing relationships on a billion-year time scale. Profile mixture models have been developed to approximate protein sequence evolution on a billion-year timescale by considering a finite mixture of stationary amino acid frequency vectors. In this talk, we address a popular concern that over-parameterization can negatively affect tree topology estimation if a large number of frequency vectors are considered. We demonstrate this is not the case via classical statistical results, extensive simulations, and empirical examples. Additionally, we introduce the GTRpmix model, implemented in IQ-TREE2, that allows maximum likelihood estimation of a common set of exchangeabilities for all site classes under any profile mixture model. We show that exchangeability matrices estimated in the presence of a site-heterogeneous profile mixture model differ markedly from those that were estimated using site-homogeneous models with a single set of equilibrium amino acid frequencies (and which are currently widely used).

Polymorphism-aware models in RevBayes: Species trees, disentangling Balancing Selection and CG-biased gene conversion

Svitlana Braichenko **1** , Rui Borges **2** , Carolin Kosiol **1,@**
1 : Centre for Biological Diversity, University of St Andrews
2 : Institut für Populatonsgenetik, Vetmeduni Vienna

The role of balancing selection is a long-standing evolutionary puzzle. Balancing selection is a crucial evolutionary process that maintains genetic variation (polymorphism) over extended periods, however, detecting it poses a significant challenge. Building upon the polymorphism-aware phylogenetic models (PoMos) framework, we introduce PoMoBalance designed to disentangle the interplay of mutation, genetic drift, directional and balancing selection pressures influencing population diversity. Rooted in the Moran model, PoMos have demonstrated efficiency

in species tree inference, capturing mutational effects, fixation biases, and GC-bias rates. Implemented in the open-source RevBayes Bayesian framework, PoMoBalance offers a versatile tool for multi-individual data analysis. This study extends PoMos' capabilities to explore balancing selection and disentangle it from GC-biased gene conversion. The novel aspect of our approach in studying balancing selection lies in PoMos' ability to account for ancestral polymorphisms and incorporate parameters that measure frequency-dependent selection. We implemented validation tests and assessed the model on the data simulated with SLiM and a custom Moran model simulator. We examined real sequences from Drosophila populations to gain insights into the evolutionary dynamics of regions subject to frequency-dependent balancing selection, particularly in the context of sex-limited colour dimorphism.