ARPIP: Ancestral sequence Reconstruction with insertions and deletions under the Poisson Indel Process

Gholamhossein Jowkar * 1,2,3 , Manuel Gil 3,4 , Julija Pecerska 3,4 , and Maria Anisimova * 3,4

1 - University of Neuchatel – Switzerland
2 - Zürich University of Applied Sciences – Switzerland
3 - Swiss Institute of Bioinformatics [Lausanne] – Switzerland
4 - Zürcher Hochschule für Angewandte Wissenschaften – Switzerland

Phylogenetics is a wide research field with a variety of applications ranging from reconstructing the tree of life to investigating ongoing epidemics. Given a phylogeny and a multiple sequence alignment from surviving species, we can infer the ancestral sequences. This insight into the evolutionary history of ancient molecules helps us understand gene function and investigate gene adaptation and convergent evolution. Here we propose a dynamic programming algorithm for joint reconstruction of ancestral sequences under the Poisson Indel Process. This modelling approach provides an explicit biological interpretation of indel events and linear time complexity for the likelihood computation with respect to the number of sequences. Our method, which we call ARPIP, consists of two steps, namely finding the most probable indel points and reconstructing ancestral sequences. First, we find the most likely indel points and prune the phylogeny to reflect the insertion and deletion events per site. Second, we infer the ancestral states on the pruned subtree in a manner similar to FastML. We applied ARPIP on a simulated dataset and on real data from the Betacoronavirus genus. We show that ARPIP reconstructs both the indel events and substitutions with a high degree of accuracy, and that our method fares well when compared to established state-of-the-art methods such as FastML and PAML. Moreover, the method can be extended to allow us to explore both optimal and suboptimal reconstructions, include rate heterogeneity through time and more. We believe it will expand the range of novel applications of ancestral sequence reconstruction.

Beyond one-gain models for pangenome evolution

Jasmine Gamblin * 1 , François Blanquart 1,2 , and Amaury Lambert 1,3

1 - Centre interdisciplinaire de recherche en biologie – Labex MemoLife, Collège de France, Centre National de la Recherche Scientifique : UMR7241, Institut National de la Santé et de la Recherche Médicale : U1050 – France
2 - Infection, Anti-microbiens, Modélisation, Evolution – Institut National de la Santé et de la Recherche Médicale : U1137, Université Paris Cité : UMR S 1 137, Université Sorbonne Paris nord − France
3 - Institut de Biologie de l'ENS Paris – Département de Biologie - ENS Paris, Institut National de la Santé et de la Recherche Médicale : U1024, Centre National de la Recherche Scientifique : UMR8197 – France

A species pangenome is the set of all genes carried by at least one representant of the species. In bacteriae, pangenomes can be much larger than the set of genes carried by one individual. Many questions remain unanswered regarding the evolutive forces shaping these bacterial pangenomes. One of them is to explain the U-shape of the gene frequency spectrum: there are more genes present in very few or almost all genomes than at intermediate frequencies. Two papers from 2012 (Baumdicker et al. and Heageman and Weitz) explained this distribution with stochastic models allowing genes to be gained only once in the species phylogeny. However the importance of intra-specific horizontal gene transfer (HGT) in many bacterial species calls for more complex models. Using a dataset of 436 commensal E.coli genomes, we show that a model with only one gain per gene is not able to reproduce the patterns of presence/absence of genes at the leaves of the phylogeny. We thus introduce a new model of pangenome evolution including a category of genes that can be gained and lost in the phylogeny multiple times, interpreted as genes undergoing frequent HGT. Both the gene frequency spectrum and the presence/absence patterns are reproduced more accurately.

Sampling consistency of diffusion statistics in Bayesian phylogeography

Pauline Rocu * 1,2,3 , Paul Bastide 3 , Denis Fargette 2 , and Stéphane Guindon 1

1 - CNRS – Laboratoire d'informatique, de robotique et de microélectronique de Montpellier (LIRMM) – France
2 - IRD – Institut de recherche pour le développement [IRD] : UMRPHIM – France
3 - CNRS – Institut Montpelliérain Alexander Grothendieck, Université de Montpellier – France

Bayesian phylogeography provides insight into the past evolution and spread of an organism using genetic and geographic data. The inference of the spatial diffusion of this organism is based on the evolution of a continuous multivariate trait: location, expressed as latitude and longitude coordinates. Common models for this continuous trait include the Brownian process (strict or relaxed (1)) and the Cauchy process (allowing jumps in the diffusion (2)).
In order to analyze and characterize the evolution, we can rely on diffusion statistics, mainly the diffusion rate (3) and the diffusion coefficient (4,5) which aim at estimating the pace at which lineages spread throughout their habitat.

However, we found that for both location models, the diffusion rate is impacted by the number of samples we are studying. More precisely, if we have N samples, the diffusion rate evolves in sqrt(N) for a Brownian process, and in log(N) for a Cauchy process. This result implies that the statistics studied would be sensitive to sampling, and that their analysis would require special consideration. Practical analyses on BEAST (6) with sequences from the West Nile Virus also showed a sampling bias on the diffusion rate value, but not in the same way as the theory predicted. Simulations were therefore carried out to test the sampling consistency of the statistics.

(1) Mandev S. Gill, Lam Si Tung Ho, Guy Baele, Philippe Lemey, and Marc A. Suchard, A Relaxed Directional RandomWalk Model for Phylogenetic Trait Evolution (2016), DOI:10.1093/sysbio/syw093
(2) Landis, Michael J et al. "Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits." Systematic biology vol. 62,2 (2013): 193-204. doi:10.1093/sysbio/sys086
(3) Nidia Sequeira Trovao, Guy Baele, Bram Vrancken, Filip Bielejec, Marc A. Suchard, Denis Fargette, and Philippe Lemey. Host ecology determines the dispersal patterns of a plant virus. Virus Evolution, 1(1):vev016, 2015. doi: 10.1093/ve/vev016.
(4) Oliver G. Pybus, Marc A. Suchard, Philippe Lemey, F. J. Bernardin, Andrew Rambaut, F. W. Crawford, R. R. Gray, N. Arinaminpathy, S. L. Stramer, M. P. Busch, and E. L. Delwart. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proceedings of the National Academy of Sciences, 109(37):15066{15071, September 2012. ISSN 0027-8424. doi: 10.1073/pnas.1206598109.
(5) Nidia Sequeira Trovao, Marc A. Suchard, Guy Baele, Marius Gilbert, and Philippe Lemey. Bayesian Inference Reveals Host-Specic Contributions to the Epidemic Expansion of Influenza A H5N1. Molecular Biology and Evolution, 32(12):3264{3275, December 2015. ISSN 0737-4038. doi: 10.1093/molbev/msv185.
(6) Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ & Rambaut A (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10 Virus Evolution 4, vey016. DOI:10.1093/ve/vey016

PPalign: Optimal alignment of Potts models representing proteins with direct coupling information

Hugo Talibart * 1 , François Coste 2 , and Mathilde Carpentier 1

1 - Institut de Systematique, Evolution, Biodiversite (ISYEB), Museum national d'Histoire naturelle, Sorbonne Universite, EPHE, UA, CNRS – Institut de Systématique, Évolution, Biodiversité, UMR 7205 ISYEB MNHN – France
2 - Univ Rennes, Inria, CNRS, IRISA – Univ Rennes, Inria, CNRS, IRISA F-35000 Rennes – France

To assign structural and functional annotations to the ever increasing amount of sequenced proteins, the main approach relies on sequence-based homology search methods, e.g. BLAST or the current state-of-the-art methods based on profile Hidden Markov Models (pHMM), which rely on significant alignments of query sequences to annotated proteins or protein families. While powerful, these approaches do not take coevolution between residues into account. Taking advantage of recent advances in the field of contact prediction, we propose here to represent proteins by Potts models, which model direct couplings between positions in addition to positional composition, and to compare proteins by aligning these models. Due to non-local dependencies, the problem of aligning Potts models is hard and remains the main computational bottleneck for their use. We introduced an Integer Linear Programming formulation of the problem and PPalign, a program based on this formulation, to compute the optimal pairwise alignment of Potts models representing proteins in tractable time. The approach was assessed with respect to a non-redundant set of reference pairwise sequence alignments from SISYPHUS benchmark which have lowest sequence identity (between 3% and 20%) and enable to build reliable Potts models for each sequence to be aligned. This experimentation confirmed that Potts models can be aligned in reasonable time (1'37" in average on these alignments). The contribution of couplings was evaluated in comparison with HHalign and independent-site PPalign. Although Potts models were not fully optimized for alignment purposes and simple gap scores were used, PPalign yielded a better mean F1 score and found significantly better alignments than HHalign and PPalign without couplings in some cases. These results, published in BMC Bioinformatics last year, show that pairwise couplings from protein Potts models can be used to improve the alignment of remotely related protein sequences in tractable time. Our experimentation suggested yet that new research on the inference of Potts models is now needed to make them more comparable and suitable for homology search. We are currently investigating in this direction, with the challenge of inferring more sensitive models with relevant coupling information.

Improving gene classification into gene families via phylo-k-mers

Benjamin Linard * 1 , Nikolai Romashchenko 1 , Vincent Lefort 1 , Emmanuel Douzery 2 , Anne-Muriel Chiffoleau 1 , Vincent Ranwez 3 , Céline Scornavacca 2 , and Fabio Pardi 1

1 - Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier – Université de Montpellier : UMR5506, Centre National de la Recherche Scientifique : UMR5506 – France

2 - Institut des Sciences de lÉvolution de Montpellier – Centre de Coopération Internationale en Recherche Agronomique pour le Développement : UMR116, Ecole Pratique des Hautes Etudes, Université de Montpellier, Institut de recherche pour le développement [IRD] : UR226, Centre National de la Recherche Scientifique : UMR5554 – France
3 - Amélioration génétique et adaptation des plantes méditerranéennes et tropicales – Centre de Coopération Internationale en Recherche Agronomique pour le Développement : UMR108, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Institut Agro - Montpellier SupAgro – France

The functional annotation of a proteome or a transcriptome is a task generally based on the classification of each of their sequence into gene families, e.g. a sets of genes that are co-orthologs from a fixed taxonomic perspective. Commonly, this task is based on similarities derived from local alignments (Blast-like approaches) Markov Model based alignments. Profiles and alignments can be built for each family independently. However, to improve the quality of the classification, one must however consider the evolutionary relationships that link the different gene families, in particular when varying evolutionary rates are characterising them. Some authors attempted to resolve this issue by computing nested profiles for different tree levels and refining the classification via phylogenetic placement tools (Tang et al, 2019, Emms et al, 2022). An alternative approach was designed by Rossier et al, with a 2-step algorithm based on k-mer indexation and tree contextualization based on ``ancestral" k-mer shared by all members of a subtree (Rossier et al, 2020). Taking into account the evolutionary relationships between families improved the classification. Such contextualization can also be performed via phylo-k-mers, a probabilistic photogenically-aware extension of the notion of k-mers that we recently developed (Linard et al, 2019, Scholz et al, 2020). From alignments and phylogenetic trees describing each gene family, phylo-k-mers can be computed. They are not necessarily observed in the input data, they are indexed when they show a high probability to have diverged from a specific branch of the family tree. I will present a new algorithm of gene classification based on phylo-k-mers. This new approach aims to reunite the power of k-mers (scalability) and phylogeny (accuracy). In particular, i) it aims to reduce the taxonomic biases encountered when profiles or alignments of strong taxonomic composition biais are used, ii) it aims to be agnostic to frame-shifts, iii) its 2-step design allows very fast classifications, and iv) optionally, it allows immediate phylogenetic placement into the family tree. We recently reinforced the scalability of this approach with the help of new algorithmic development dedicated to the production of fast and efficient phylo-k-mer computations. I will briefly discuss these developments.

References :

Emms et al. SHOOT: phylogenetic gene search and ortholog inference. Genome Biol 23, 85 (2022).
Linard et al. Rapid Alignment-free Phylogenetic Placement via Ancestral Sequences. Bioinformatics, Volume 35, Issue 18, 15 September 2019, Pages 3303–3312.
Rossier et al. OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches, Bioinformatics, Volume 37, Issue 18, 15 September 2021, Pages 2866–2873.
Scholz et al. Rapid screening and detection of inter-type viral recombinants using phylo-k-mers, Bioinformatics, Volume 36, Issue 22-23, 1 December 2020, Pages 5351–5360.
Tang H. et al. (2019) TreeGrafter: phylogenetic tree-based annotation of proteins with Gene Ontology terms and other annotations. Bioinformatics, 35, 518–520.

Mutual Information-based Feature Selection of Informative Phylo-k-mers

Nikolai Romashchenko * 1 , Benjamin Linard 2,1 , Fabio Pardi 1 , and Eric Rivals 1

1 - LIRMM – Centre National de la Recherche Scientifique : UMR5506, Université de Montpellier : UMR5506 – France
2 - Spygen – SPYGEN [Le Bourget-du-Lac] – France

Phylo-k-mers is a probabilistic phylogenetically-aware extension of the notion of k-mers. For a reference alignment of sequences of a genomic region and a phylogenetic model of sequence evolution, they describe what k-mers can be observed at different locations of a fixed phylogeny and with what probability. Computation of phylo-k-mers is a prerequisite for the recently proposed methods of alignment-free phylogenetic placement and detection of novel viral recombinants. Many k-mers - potentially 4ˆk for DNA and 20ˆk for protein sequences - need to be considered at each tree node, making the resulting collections of phylo-k-mers large in size.

We proposed an information-based method of selecting phylo-k-mers that are informative for phylogenetic placement. The method adapts an existing feature selection approach for text classification, computing Mutual Information between the tree branch variable and the variable indicating the presence of a k-mer in hypothetical sequences originating from this branch. Experiments on phylogenetic placement using filtered sets of phylo-k-mers showed that accurate phylogenetic placement can be performed using only small fractions of the most informative phylo-k-mers: 6% of their total number for placing rbcL gene sequences and 12% for 16S rRNA gene sequences.

Preprocessing Strategies for Bayesian Phylogeographic Analysis Using Large-Scale Genomic Sequence Data

Yimin Li ∗ 1 , Augustin Clessin 1,2 , Samuel Hong 1 , Nena Bollen 1 , and Guy Baele 1

1 - Department of Microbiology, Immunology and Transplantation [Leuven] – Belgium
2 - École Normale Supérieure de Lyon, Université Claude Bernard Lyon 1 – Université de Lyon, Université Lyon 1 – France

The ongoing SARS-CoV-2 pandemic has been posing a huge threat to public health, economic development and social interactions since the end of 2019. Different SARS-CoV-2 variants keep emerging throughout this pandemic and are important to study in terms of their evolution, local and/or global dispersal, impact on transmissibility, severity, and immunity. First detected in December 2020, SARS-CoV-2 lineage B.1.525 contains several mutations of biological significance. The E484K mutation and ΔY144 deletion tend to drive immune escape, while the D614G mutation and ΔH69/V70 deletion can increase transmissibility and infectivity. With nearly ten thousand genomes from this lineage being available, conducting a detailed Bayesian phylogeographic analysis on the complete data is not feasible. We explore different strategies for reducing this data set to a representative set of genomic sequences that enable us to reconstruct the origin and dispersal history of this lineage. Initial data exploration strategies such as provided in TempEst have yielded different results depending on data set size, with the complete data set seemingly evolving at half the evolutionary rate as a data set that consists only of high-quality genomes. We first explore various maximum-likelihood and Bayesian inference methodologies - paying special attention to different molecular clock and tree prior specifications - on this core high-quality data set to establish a consensus for both TMRCA and mean evolutionary rate, which we compare to estimates from other SARS-CoV-2 lineages. We subsequently evaluate the temporal signal in the remaining genomes, grouping by time, sequencing lab and country, to determine which genomes bias the temporal signal in the core data set and warrant further investigation. To this end, we use several popular inference packages, such as BEAST, TreeTime and Chronumental. When the final data set has been constructed, we employ several subsampling procedures to avoid sampling bias, as this might impact estimation of important phylogenetic and phylogeographic parameters. Targeted at analysing thousands of viral sequences, our work aims to provide a reproducible genomic data (pre-)processing pipeline for (SARS-CoV-2) phylogeographic inference analyses.

Combinatorics of multiple-merger coalescent genealogies

Johannes Wirtz ∗ 1

1 - Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier – Centre National de
la Recherche Scientifique : UMR5506 – France

Sample genealogies generated according to Kingman's Coalescent are always binary. More precisely, they are elements of the class of binary time-labelled trees of finite size. The combinatorial aspects of this class are well-understood, and it is also well-known what probability distribution is imposed on them by the process. However, under a lot of other population models, such as the Lambda-Coalescents, of which Kingman's Coalescent is a special case, or various spatial models (e.g. the spatial Lambda-Fleming-Viot, Branching Brownian Motions etc), internal nodes of sample genealogies do not need to be binary. The combinatorial class of general time-labelled trees is not as well-understood as their binary counterpart, and there are also few results regarding what probability distribution is induced on them by the differend models. We make use of techniques borrowed from the field of analytic combinatorics to study the asymptotics of the number of trees of any size, and point out some relations between the Lambda-Coalescents (particularly, the so-called Psi-Coalescents) and their respective tree probability distributions.

Leveraging tools from Nextstrain for bespoke phylogenetic analysis of viral pathogen epidemics

Barney I. Potter ∗ 1 , John Huddleston 2,3 , Cornelius Roemer 4,5 , Elias Harkins 2 , Ivan Aksamentov 4,5 , James Hadfield 2 , Jennifer Chang 2 , Jover Lee 2 , Kairsten Fay 2 , Sidney M. Bell 6 , Thomas R. Sibley 2 , Victor Lin 2 , Samuel L. Hong 7 , Emma B. Hodcroft 5,8 , Richard A. Neher 4,5 , Trevor Bedford 2,3,9 , and Guy Baele 1

1 - Rega Institute - KU Leuven – Belgium
2 - Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center – United States
3 - Molecular and Cellular Biology Program, University of Washington – United States
4 - Biozentrum, University of Basel – Switzerland
5 - Swiss Institute of Bioinformatics – Switzerland
6 - Chan Zuckerberg Initiative – United States
7 - Rega Institute - KU Leuven – Belgium
8 - Institute of Social and Preventive Medicine, University of Bern – Switzerland
9 - Howard Hughes Medical Institute – United States

Timely, actionable analysis of emergent viral outbreaks is vital to inform effective public health intervention. Phylogenetic and phylogeographic techniques represent powerful, flexible tools for inference of epidemiological factors of a viral outbreak that are of public health import: reproductive number, evolutionary rate, effective viral population size, and the timing of geographic migrations can all be inferred using phylogenetic techniques. Today, a key challenge to using these methods for the analysis of viruses of great threat to human health-particularly SARS-CoV-2, the causative agent of COVID-19-is the vast quantity of genomic sequence data and associated metadata that are generated every day. Due to the computational complexity of phylogenetic modeling techniques, it is necessary for researchers to employ principled methods of data selection and analysis to strike a balance between ensuring analytical timeliness through subsampling, thereby preserving key trends that may inform public health interventions. Nextstrain, a phylogenetics toolkit that facilitates analysis from database-to-visualization, sits at the forefront of real-time phylogenetic and genomic epidemiological analysis. Here, we describe a methodology by which Nextstrain tools can be used in conjunction with other existing phylogenetic analysis software and custom scripts to create bespoke analyses of SARS-CoV-2 genomic sequence data. Our novel methodology supplements existing pipelines in three major ways. First, we show how high-resolution geographic metadata may be incorporated into existing Nextstrain workflows to facilitate the identification and characterization of local viral transmission chains within the context of a single country's SARS-CoV-2 landscape. Second, we enable and showcase the use of custom phylogenetic inference approaches within Nextstrain for tree building, and ancestral state reconstruction. Finally, we demonstrate how Nextstrain tools used in conjunction with the novel methodologies we describe can be used to inform further analyses that account for phylogenetic and phylogeographic uncertainty.

Modeling the dynamics of antibiotic resistance genes: towards an ecology of the bacterial pangenome

Rémi Tuffet * 1 , Gabriel Carvalho , Anne-Sophie Godeux , Maria-Halima Laaberki , Samuel Venner , and Xavier Charpentier

1 - Université Lyon 1 – UMR5558 LBBE, CIRI, Inserm, U1111, Université Claude Bernard Lyon 1, CNRS, UMR5308, École Normale Supérieure de Lyon, Univ Lyon, F-69007, Lyon, France – France

One of the major challenges for human and animal health in the coming decades is to control the emergence and spread of antibiotic resistance (AR) in bacteria. Mobile genetic elements (MGEs) are the main vehicles of AR genes and their evolutionary trajectories are partly independent of those of their bacterial hosts. Understanding the dynamics of AR requires understanding the co-evolution of bacterial and MGE strategies, and identifying the forms of evolutionary cooperation and conflict between these entities. Here, we focus on the evolution of MGE strategies in the bacterium Acinetobacter baumannii (Ab). Ab represents a growing public health problem due to its level of resistance to antibiotics. In Ab, as in many other bacteria, natural transformation (acquisition of extracellular DNA controlled by bacteria) is a way to cope with environmental stochasticity by ensuring both the acquisition of new genes (e.g. from MGEs carrying AR genes) and their elimination. While MGEs carrying AR genes could insert themselves randomly into the genome, we showed that in more than 96% of the cases, they integrate into a specific gene (comM) which results in a drastic (but not total) reduction of the transformation rate of bacteria. This suggests that MGEs cooperate with bacteria (by carrying AR genes) but also have a form of conflict with them (by partially inhibiting natural transformation). To test this hypothesis, we propose a model to quantify the success of competing MGEs while inserting into sites that partially, not at all or completely inhibit transformation. From this model parameterized with experimental data, we observed that the insertion of MGEs into the site that partially inhibits natural transformation confers a selective advantage to MGEs when the host bacteria are exposed to a stochastic environment. This strategy allows MGEs to minimize the risk of being eliminated from genomes by natural transformation, while ensuring that their host cells maintain a basal transformation activity that allows them to acquire other adaptive genetic determinants, thus favoring the propagation of these MGEs. This work shows that the probability of persistence of AR genes depends strongly on the strategies of the MGEs that carry them. It is therefore necessary to develop new approaches to better understand the eco-evolutionary dynamics of MGEs and bacteria (alternating conflicts and evolutionary cooperation) to understand the dynamics of the AR and more broadly the dynamics of bacterial pangenomes. We propose to develop new models inspired by theories and tools of ecology to understand this complexity, going beyond the classical approaches of genomics.

Estimation of reproductive number and prevalence using genomic and time series data

Alexander Zarebski * 1 , Louis Du Plessis 2 , Kris Parag 3 , and Oliver Pybus 1,4

1 - University of Oxford – United Kingdom
2 - ETH Zurich – Switzerland
3 - Imperial College London – United Kingdom
4 - The Royal Veterinary College – United Kingdom

In genetic epidemiology, current methods can struggle with the size of pathogen genome datasets. While the methodology for processing increasingly large datasets receives substantial attention, less progress has been made in integrating additional types of data into analyses. For example, utilising both pathogen genomes and time series of confirmed cases to inform estimates. Mechanistic models of transmission and observation (e.g. the birth-death-sampling model) provide a natural approach to include additional streams of data but can be challenging to apply in practice. To utilise both genomic and case data on a large scale, we developed an efficient and accurate approximation scheme, called TimTam, capable of estimating both the reproductive number and the prevalence of infection. This method (with piecewise constant rate parameters) will soon be available as a BEAST2 package.

Bridging the gap between population genomic and phylogenetic approaches

Mélodie Bastian * 1 and Nicolas Lartillot 2

1 - Bioinformatique, phylogénie et génomique évolutive – Département PEGASE [LBBE] – France
2 - Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Université de Lyon, Institut National de Recherche en Informatique et en Automatique, VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Centre National de la Recherche Scientifique : UMR5558 – France

The availability of an increasing amount of genetic data has resulted in a boom in molecular evolution studies in recent years and a better elucidation of evolutionary mechanisms such as genetic drift. The intensity of the genetic drift is inversely proportional to the number of breeders in the population, i.e. the effective size (Ne). It can impact both short and long term evolutionary processes (estimated respectively from polymorphism and divergence data). The aim of my work is to contrast intraspecific and interspecific data in order to bridge the gap between population genomic and phylogenetic analyses.

In this study, I estimate variations in Ne between species based on genome-wide heterozygosity, correcting for variations in mutation rate (μ), along the mammalian phylogeny in order to study correlations between Ne, ecological traits and molecular traits such as selection intensity (both long term, based on dN/dS, and short-term, based on piN/piS). For this purpose, I devised a pipeline from the recovery of orthologous gene sequences, ecological and heterozygosity data to Bayesian integrative analysis aiming at reconstructing Ne by a multivariate process (1). I observe that the Ne of very massive animals is much smaller than the Ne of small mammals. I obtain positive corelations between dN/dS and life history traits, consistent with previous analyses (2) and suggestive of a role of Ne. The more direct correlation of traits with heterozygosity and Ne is still under investigation.

1. Brevet, M. & Lartillot, N. Reconstructing the history of variation in effective population size along phylogenies. 793059 https://www.biorxiv.org/content/10.1101/793059v4 (2021) doi:10.1101/793059.
2. Figuet, E. et al. Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes. Mol Biol Evol 33, 1517–1527 (2016).

Convolutional graph networks for Coevolution detection in COVID19

David Moi * 1,2 and Christophe Dessimoz 1,2

1- University of Lausanne – Switzerland
2- Swiss Institute of Bioinformatics [Lausanne] – Switzerland

Since the start of the COVID19 pandemic we have had access to an unprecedented volume of data available to inform epidemiological and healthcare policy decisions, but making sense of this deluge of genomic data has been difficult using current bioinformatics methods. One technique which may provide insight into COVID's evolutionary trajectory in response to the shifting fitness landscape is to study coevolutionary patterns in the genome or protein sequences. Finding these evolutionary associations could point to shared constraints in the viral replication cycle between multiple sites and provide valuable targets for treatment and prevention. Unfortunately, finding coevolutionary signals between sites of the viral genome or the viral proteome has so far been prohibitively computationally expensive due to quadratic

time complexity associated with comparing all 30K sites combined with the sheer number of sequences (over 3.5 million and rapidly increasing). This presentation showcases tools built around distributed computing, probabilistic data structures and machine learning to deal with these problems and extract pairs of coevolving sites from the GISAID database containing all available COVID sequences. While the COVID19 pandemic is currently unique in its density of sampling, we may see similar datasets emerge in coming years in the context of continuous metagenomic monitoring, in the unfortunate event of further pandemics or in other heavily sequenced species such as humans. We hope that the ideas presented in this work will also see adoption in these emerging areas.

Bayesian model comparison of molecular clock models - a phylogenetic simulation study

Kanika Nahata * 1 , Mandev S. Gill , Karthik Gangavarapu , Marc Suchard 2,3,4 , and Guy Baele 5

1 - Rega Institute – Belgium
2 - Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA – United States
3 - Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA – United States
4 - Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA – United States
5 - Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven – Belgium

In the 1960s, several groups of scientists - including Emile Zuckerkandl, Linus Pauling and Allan Wilson - had noted that proteins experience amino acid replacements at a surprisingly consistent rate across very different species. Since the proposal of such a (strict) clock model, a wide range of different clock model parameterizations have emerged which now take up a prominent place in the field of phylogenetic inference as well as in many other areas of evolutionary biology. In studying pathogen evolution, molecular clocks allow combining the genetic differences between samples and their collection times to estimate time-calibrated phylogenies. Along with the development of increasingly complex clock models comes the need to accurately determine which model is best suited to analyse a particular data set. For this purpose, different marginal likelihood estimators have been developed in recent years to compare relative model fit in a Bayesian framework. These estimators have shown considerable improvements in accuracy, but often at the expense of an increased computational cost. In our simulation study, we examine the performance of these estimators in identifying the correct underlying molecular clock model.

Protein folds as synapomorphies of the tree of life

Martin Romei 1 , Guillaume Sapriel 1 , Jacques Chomilier 2 , Guillaume Lecointre 1 , and Mathilde Carpentier * 1

1 - Institut de Systématique, Evolution, Biodiversité – Museum National d'Histoire Naturelle, Ecole Pratique des Hautes Etudes, Sorbonne Universite, Centre National de la Recherche Scientifique : UMR7205, Université des Antilles – France
2 - Institut de minéralogie, de physique des matériaux et de cosmochimie (IMPMC) – Museum National d'Histoire Naturelle, Institut de recherche pour le développement [IRD] : UR206, Sorbonne Universite : UM120, Centre National de la Recherche Scientifique : UMR7590 – Tour 23 - Barre 22-23 - 4e étage - BC 115 4 place Jussieu 75252 PARIS, France

Structural domains of proteins are defined by the connectivity and organisation in three dimensions of their secondary structures: the ``fold". The total number of folds is quite stable, about 1200 which is surprisingly low. It is possible that folds are extremely reliable characters, thus reliable phylogenetic entities, complementary to traditional phylogenetic signals. Moreover, their slower dynamic of change could allow to highlight deep evolution of organisms. We have explored the repartition of folds within the tree of Life to evaluate their potential as phylogenetic markers and trying to answer the question of structural convergence by measuring the consistency of their repartition within a reference phylogeny. We mapped folds onto a tree of life and measure the consistency of each fold character. We have developed and explored a methodology which allows us to analyze the repartition of folds relying on a seriated heatmap and several clusterings.

Our results show that 20% of the folds are present in all superkingdoms, and 53.9% are potential synapomorphies. We find fold characters consistently supporting several nested eukaryotic clades with divergence times spanning from 1,100 mya to 380 mya. As for the earliest branches of the tree of life, the three superkingdoms are discriminated by eukaryotic specific folds (181) as well as shared folds between Eukaryota and one of the two other superkingdoms. Many folds shared by parts of eukaryotes and some eubacteria should result from past horizontal transfers (e.g. cyanobacteria to photosynthetic eukaryotes) witnessing significant fold flow to eukaryotes. Among eukaryotes, some folds therefore appear as synapomorphies of the species phylogeny, while others are markers of transfers to Eukaryota.

We have highlighted that folds are reliable synapomorphies. They are witnesses of ancient events like primary and secondary endosymbiosis, but they can be specific of more recent clades like metazoan or vertebrates. We have also analyzed functions of folds inherited from archaea and bacteria which reveal in both cases an over-representation of informational function. For specific eukaryote folds, we observe an overrepresentation of regulation functions linked to extra-cellular mechanism matching with multicellularity appearance.

Folds provide information for reconstructing the history of life as witnesses of major evolutionary events or as witnesses of the appearance of new functions during evolution. They can also provide valuable insights into the evolution and design of protein folds: how do they arise, evolve and adapt to specific functions? However, our work highlights the complex history of folds and the need to untangle the evolutionary events of lateral transfers to understand the divergence and apparition of new protein structure. Here we would like to discuss with the community the best approaches to address these questions.

Evaluation of methods to detect shifts in directional selection at the genome scale

Louis Duchemin, Philippe Veber, and Bastien Boussau ∗ 1

1 - Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Université de Lyon, Centre National de la Recherche Scientifique : UMR5558 – 43 Bld du 11 Novembre 1918 69622 VILLEURBANNE CEDEX, France

Identifying the footprints of selection in coding sequences can inform about the importance and function of individual sites. Analyses of the ratio of non-synonymous to synonymous substitutions (dN /dS) have been widely used to pinpoint changes in the intensity of selection, but cannot distinguish them from changes in the direction of selection, i.e., changes in the fitness of specific amino acids at a given position. A few methods that rely on amino acid profiles to detect changes in directional selection have been designed, but their performance have not been well characterized. In this paper, we investigate the performance of 6 of these methods. We evaluate them on simulations along empirical phylogenies in which transition events have been annotated, and compare their ability to detect sites that have undergone changes in the direction or intensity of selection to that of a widely used dN /dS approach, codeml's branch-site model A. We show that all methods have reduced performance in the presence of biased gene conversion but not CpG hypermutability. The best profile method, Pelican, a new implementation of (Tamuri et al., 2009), performs as well as codeml in a range of conditions except for detecting relaxations of selection, and performs better when tree length increases, or in the presence of persistent positive selection. It is fast, enabling genome-scale searches for site-wise changes in the direction of selection associated with phenotypic changes.

Uncovering the Diverse Roles of Short Tandem Repeat Variation in Colorectal Cancer

Max Verbiest ∗ 1,2,3 , Oxana Lundström 1,4 , and Maria Anisimova 1,3

1 - Zurich University of Applied Sciences – Switzerland
2 - University of Zurich – Switzerland
3 - Swiss Institute of Bioinformatics – Switzerland
4 - Stockholm University – Sweden

Short tandem repeats (STRs, also known as microsatellites) are highly variable, back-to-back repetitions of small DNA motifs. Insertions and deletions of repeat motifs are common in STR loci and can affect gene expression levels and protein structures. Their high mutability can result in tumor-specific neoantigens, making them promising targets for cancer vaccines and immune therapies. Although recently developed computational approaches allow for accurate genotyping of STRs from sequencing data, many investigations of STR variation in cancer predate these methods. We therefore suspect that the contribution STRs have to the molecular picture of cancer is currently underestimated. To investigate this, we used repeat-specific methods to generate and genotype a panel of over 1.8 million STR loci in colorectal cancer (CRC) patients from The Cancer Genome Atlas. We stored both the panel of STR loci and their variability in CRC patients in a PostgreSQL database, which will be made available to the scientific community through a web interface for easy access. We detected tumor STR variants by comparing repeat lengths between patient-matched healthy and diseased tissue. We then estimated the contribution of these tumor STR variants to gene expression changes in CRC using existing catalogues of STRs known to affect expression. For STR variants in coding regions, we determined the expected changes in protein structure and monitored the potential generation of targetable neoantigens. While a lot of this is still work in progress, we expect our results will provide a better understanding of the diverse roles of STR variation in CRC. Using computational methods specifically designed to analyse STRs, we will demonstrate the importance of this abundant but often bypassed source of variation in cancer. Furthermore, by making our panel of STR loci and their variation in CRC patients available to the community, we hope to stimulate future investigations into this important topic.

The influence of genetic dosage on PRDM9-dependent evolutionary dynamics of meiotic recombination

Alice Genestier * 1 , Nicolas Lartillot 2 , and Laurent Duret 3

1 - Département PEGASE [LBBE] – Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – France
2 - Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Université de Lyon, Institut National de Recherche en Informatique et en Automatique, VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Centre National de la Recherche Scientifique : UMR5558 – France
3 - Laboratoire de Biométrie et Biologie Evolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 Bld du 11 Novembre 1918 69622 VILLEURBANNE CEDEX, France

Meiosis is an important step in the eukaryotic life cycle during which recombination and proper chromosome segregation takes place. In mammals, recombination is regulated by the Prdm9 gene. This gene, which possesses a double function (recruitment of the double strand break machinery and facilitation of the pairing of homologous chromosomes), induces an intra-genomic Red Queen resulting from the opposition of two antagonistic forces : erosion of the recombination landscape by biased gene conversion and positive selection on PRDM9. This Red Queen was previously modeled, but without taking into account the role of PRDM9 as a pairing facilitator. Accordingly, I developed a mechanistic model taking into account the dual role of PRDM9. This modeling work gives important insights into the Red Queen mechanism, thus completing previous studies. In particular, it reveals that positive selection of new PRDM9 alleles is due to the reduced symmetrical binding caused by the loss of high affinity binding sites and, on the other hand, it demonstrate the influence of the genetic dosage of PRDM9 on the dynamics of the Red Queen, which can result in negative selection on new PRDM9 alleles entering the population.

An Individual-based model to study the importance of trade-offs in the evolution and diversification of traits in host phage population dynamics and long-term co-existence

Fateme Pourhasanzade * 1 , Swami Iyer 2 , and Selina Vage 1

1 - University of Bergen – Norway
2 - University of Massachusetts – United States

It has revealed over the last few decades that viruses play an important role in the evolution of all species, shaping populations and biodiversity in marine ecosystems. Mathematical and agent-based models are useful tools in studying/predicting the surprising patterns in virus-host dynamics, especially when we consider the fact that monitoring viral population parameters are challenging in the field and laboratory. We have developed an individual-based model to explore the dynamics of host and virus populations. Host dynamics are validated with lab results for different initial multiplicities of infection (MOI). We have studied the impact of coevolution and showed the importance of trade-offs between competitive and defensive host traits that can shape biological interactions and diversity in the host-phage dynamics. We have also investigated the effect of several organismal and environmental parameters such as the burst size of viruses on the dynamics of host and virus populations. Our model serves as a powerful tool to study bacteria phage interactions in different environmental settings.

GPU-accelerated online phylodynamic inference using BEAST

Samuel Hong * 1 , Philippe Lemey 1 , and Guy Baele 1

1 - Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven – Belgium

The COVID-19 pandemic has put genomic epidemiology at the forefront of the pandemic response, with genome sequencing as an essential tool for tracking the emergence and spread of novel SARS-CoV-2 variants. Within this context, Bayesian phylodynamics provides a powerful framework to jointly reconstruct the evolution and spread of new variants while incorporating non-genomic data such as travel history and mobility information. A limitation of Bayesian phylodynamic analyses is that they take considerable time to run, due to the fact that sampling trees through Markov chain Monte Carlo (MCMC) is a very time-consuming process. This is specially cumbersome during an ongoing epidemic as continuous data generation requires frequently updated analyses. To account for this, a data augmentation procedure had previously been implemented in BEAST, allowing for phylodynamic inference in an online fashion. This approach has been shown to significantly reduce the burn-in time required for MCMC to converge. In addition, BEAST - in conjunction with the BEAGLE library - allows for GPU acceleration to further reduce the required computation time. In this study, we aim to assess the performance gains that can be obtained by parallelizing BEAST runs using multiple GPUs when performing online inference. For this purpose, we considered an early epidemic scenario by constructing an initial alignment of 502 SARS-CoV-2 genomes sampled during the beginning of the pandemic, which we updated 10 times at equally spaced time intervals using the online inference implementation in BEAST. At each time

point, we ran 8 parallel GPU accelerated analyses and calculated the time required to obtain an effective sample size (ESS) of 200 when combining results using 1 to 8 GPUs. Our results show that using multiple GPUs significantly reduces the runtime required to obtain ESS > 200, with up to a 7.4-fold reduction in runtime across all dataset sizes (n=502 to 2,108) using 8 GPUs. Additionally, we observe that the marginal reduction in runtime decreases as more GPUs are included in the analysis, with 6 GPUs being the threshold after which we start observing diminishing returns.

Intergeneric relationships within Ophioglossaceae untangled with organelle phylogenomics

Darina Koubinova * 1

1 - Université de Neuchâtel – Switzerland

Ophioglossaceae is one of fern families comprising of several ancient-diverged lineages, and some ancient lineages in this family contain only few living remnants, such as the monotypic subfamilies Helminthostachyoideae and Mankyuoideae. The four earliest diverged lineages among extant Ophioglossaceae are now recognized as different subfamilies, but their relationships remain unresolved seemingly due to insufficient phylogenetic efforts in previous studies. Former attempts to infer phylogenetic tree structures included only limited plastid regions, and some of them even contained a great portion of missing data. In two one phylogenomic analyses, the plastome dataset was used but scarce representatives with only one or two for each subfamily (or only three of them) sampled. Besides, a rather simplified substitution model applied in these phylogenomic datasets might introduce systematic errors. To tackle the deepest and also difficult nodes in Ophioglossaceae, we adopted a phylogenomic approach with adding more subfamily representatives (9 out of total of 12 currently recognized genera) and analyzed datasets from not only plastome but also mitogenome. We used genome skimming data to assemble these organelle genomes and from the resulting assemblies, we extracted the coding sequences (CDS) for the phylogenomic inferences. We tested different partition and substitution models for these phylogenomic datasets, including finer ones in order to better account for rate heterogeneity among loci and codon positions. Our phylogemonic results overall supported a novel, previously uncovered topology which presented the most solid infra-family backbone for Ophioglossaceae. Finally, based on this infra-family backbone, we traced phylogenetic origins of the hypothesized horizontal gene transfer (HGT) in organellar genomes, ancient whole genome duplication (WGD) events, and key morphological innovations in Ophioglossaceae.

The robustness of bootstrap branch supports with respect to taxon sampling

Paul Zaharias 1, Frédéric Lemoine 2 & Olivier Gascuel 1

[1]Institut de Systématique, Évolution, Biodiversité (UMR7205 - MNHN, CNRS, SU, EPHE, UA), Paris, France
[2]Institut Pasteur, Université de Paris, Paris, France

The bootstrap method is based on resampling alignments and reestimating trees. Felsenstein's bootstrap proportions (FBP; Felsenstein 1985) is the most common approach to assess the reliability and robustness of sequence-based phylogenies. However, when increasing taxon-sampling (i.e., the number of sequences) to hundreds or thousands of taxa, FBP will tend to return low supports for deep branches. The Transfer Bootstrap Expectation (TBE; Lemoine et al. 2018) has been recently suggested as an alternative to FBP. TBE is measured using a continuous transfer index in [0,1] for each bootstrap tree, instead of the {0,1} index used in FBP to measure the presence/absence of the branch of interest. TBE has been shown to yield higher and more informative supports, without inducing falsely supported branches. Nonetheless, it has been argued that TBE must be used with care due to sampling issues, especially in datasets with high number of closely related taxa. In this study, we conduct multiple experiments by varying taxon sampling and comparing FBP and TBE support values on different phylogenetic depth, using both simulated and empirical datasets. Our results show that the main critic of TBE stands in extreme cases, but that TBE is still very robust to taxon sampling in most simulated and empirical cases, while FBP is inescapably negatively impacted by high taxon sampling. We suggest guidelines and good practices in TBE computing and interpretation.

Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap, *Evolution*, *39*(4):783-791, 1985.
Lemoine F, Domelevo-Entfellner JB, Wilkinson E, Correia D, Dávila Felipe M, De Oliveira T, Gascuel O. Renewing Felsenstein's phylogenetic bootstrap in the era of big data, *Nature*, 556(7702): 452-456, 2018.

The Ploidy Level of Phylogenetic Networks

Katharina T Huber, Liam J Maher

University of East Anglia

Polyploidy is an important evolutionary condition affecting organisms ranging from animals to plants. It is observed in the data as multiple complete sets of chromosomes, or ploidy level. Reconstructing the evolutionary past of polyploid organisms is a complex task.